Augusto Lucubo

Licenciado em Matemática

Seguro de Responsabilidade Civil Automóvel
Modelos de Tarifação

Dissertação para obtenção do Grau de Mestre em

Actuariado, Estatística e Investigação Operacional.

Orientadora: Professora Doutora Gracinda R. Guerreiro,
Auxiliar Professor,
Universidade Nova de Lisboa

Júri

Presidente:	Maria Isabel Azevedo Rodrigues Gomes
Arguente:	Manuel Leote Tavares Inglês Esquível
Vogal:	Gracinda Rita Diogo Guerreiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março, 2020

Seguro de Responsabilidade Civil Automóvel Modelos de Tarifação

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

"A persistência leva-nos a alcançar a meta trançada."
-Augusto Lucubo

Agradecimentos

Á Deus Todo-Poderoso pela oportunidade de alcançar este nível, pelo sustento em momentos difíceis e a Igreja Cristã Unida de Lisboa, em particular a irmã Glória Manjinela pelo apoio espiritual.

Á Professora Gracinda Guerreira, orientadora desta dissertação, pela oportunidade em trabalho comigo. Á todos os professores do curso de Matemática e Aplicações, especialmente aos do ramo Actuariado, Estatística e Investigação Operacional, pelos conhecimentos transmitidos.

Áo Dr. Henda Mondlane Ferreira da Silva pela cedência de dados que suportaram desta dissertação e o Dr. Gil Wilson Morais pela disponibilidade.

Agradeço a minha querida esposa **Juliana Lucubo**, que mais sofreu pela minha ausência, tornando-se mãe e pai aos mesmo tempo, cuidado da Nilce Lucubo, Yannis Lucubo, Yohanna Lucubo e a Chelsia Domingos, que também agradeço.

Áos padrinhos Dikila e Mário Silva pelo apoio incondicional.

Áos meus pais João Ntela e Dina Paulina e os meus irmãos.

Á todos que direta ou indiretamente participaram no alcançar deste objectivo.

Resumo

O aumento do número de empresas de Seguros em Angola, consequência da obrigatoriedade de seguro de Responsabilidade Civil Automóvel, impõe competitividade no sector e um nível concorrencial elevado na venda dos produtos. Por esta razão, a diferença observada entre os produtos comercializados por diferentes empresas pode reflectir-se na decisão do segurado. Neste contexto, é necessário a definição de uma tarifa tecnicamente equilibrada que permita à empresa assegurar o cumprimento das suas responsabilidades, mas que seja também justa e adaptada a cada cliente. A técnica a estudar neste trabalho leva a identificar, diferenciar e quantificar o grau de risco de cada segurado permitindo, assim, cobrar o prémio adequado e não incorrer em perdas financeiras.

O objecto deste trabalho consiste em levar a cabo uma análise estatística de Frequência e Severidade da Sinistralidade no Seguro de Responsabilidade Civil Automóvel, a qual permitirá à seguradora oferecer soluções técnicas a problemas de tarifação, consequentemente o cálculo do prémio, mediante um processo no qual se deve ter em conta os aspectos técnicos, reguladores, económicos e estatísticos, tendo por base dados reais de uma seguradora Angolana.

A metodologia aplicada incidirá sobre Modelos Lineares Generalizados (MLG), partir da qual se desenvolve um processo de modelação sobre um conjunto de dados de uma carteira de Responsabilidade Civil Automóvel, visando a construção de tarifa *a priori*.

Palavras-chave: Tarifação, Seguro Automóvel, Modelos Lineares Generalizados, Família Exponencial.

Abstract

The increase in the number of insurance companies in Angola, as a consequence of the mandatory liability insurance, imposes competitiveness in the sector and a high level of competition in the sale of products. For this reason, the difference observed between the products marketed by different companies may be reflected in the insured's decision. In this context, it is necessary to define a technically balanced tariff that allows the company to ensure the fulfillment of its responsibilities, but that is also fair and adapted to each customer. The technique to be studied in this work leads to the identification, differentiation and quantification of the degree of risk of each insured, thus allowing the collection of the appropriate premium and not incurring financial losses.

The object of this work is to carry out a statistical analysis of Frequency and Severity of claims in the Automobile Liability Insurance, which will allow the insurer to offer technical solutions to pricing problems, consequently the premium calculation, through a process in which must take into account the technical, regulatory, economic and statistical aspects, based on real data from an Angolan insurer.

The applied methodology will focus on Generalized Linear Models (GLM), from which a modeling process is developed on a set of data from an Automobile Liability portfolio, aiming at the construction of a priori tariff.

Keywords: Pricing, Auto Insurance, Generalized Linear Models, Exponential Family.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O objectivo principal das Companhias de Seguros consiste em gerenciar riscos transferidos pelo segurado, mediante as condições estabelecidas na apólice de seguro, em que irá indemnizar o segurado na ocorrência de sinistros. Para tal, a seguradora estabelece o prémio a pagar pelo segurado, tendo em conta o risco que possui. No Seguro Automóvel, a sinistralidade é influenciada por factores endógenos e exógenos do condutor e do automóvel, considerados na altura do processo de tarifação, e recolhidos para cada cliente durante o processo de subscrição, pois são determinantes na estimação do prémio para cada apólice.

A tarifação do Seguro Automóvel, de um ponto de vista técnico, tem um papel fundamental na sustentabilidade das Companhias de Seguros, pois tem como objectivo principal o correcto cálculo de prémio de forma equitativa, justa e suficiente, tendo em conta o risco que incorre o tomador de seguro e ainda o pagamento de um prémio ajustado ao risco.

A ciência actuarial apresenta ferramentas necessárias e suficientes, oferecendo soluções técnicas a problemas de tarifação, isto é, sugere modelos de cálculo do prémio, justificáveis teoricamente e com alto grau de confiança. Dai a importância de levar a cabo uma análise estatística da **Frequência de Sinistralidade** e **Severidade de sinistro**, baseando-se em princípios estatísticos, como:

- A Estatística, ciência que procura reunir informações quantitativas e qualitativas referentes a indivíduos, grupos, séries de eventos, que realiza a análise desses dados, incluindo a proporção ou ocorrência de sucessos e fracassos de

um evento, e deles derivam significados precisos ou previsões para o futuro.

- Probabilidades, ciência que fornece uma explicação matemática a resultados que aparecem em experiências ou acontecimentos de risco.
- A Lei dos Grandes Números afirma que à medida que o número de observações independentes de uma população aumenta, a média da amostra aproxima-se da média da população cada vez mais, permitindo obter probabilidades mais precisas sobre perdas da seguradora

A abordagem deste trabalho consiste em construir uma tarifa de Responsabilidade Civil Automóvel de uma empresa de seguros, o mais adequada possível às características do cliente, isto é, encontrar um modelo adequado para a frequência e severidade de sinistro de tal modo que se consiga prever o número esperado de sinistros por unidade de tempo para um indivíduo com determinado perfil, assim como o custo esperado por sinistro que esse indivíduo possa vir a custar à Companhia, com base nos Modelos Lineares Generalizados, que permitem estimar os efeitos de um determinado factor nas observações.

O presente trabalho é composto por uma introdução e quatro capítulos. No segundo capítulo são tecidas considerações essenciais para o entendimento dos modelos de tarifação. O terceiro capítulo reserva-se a revisão literária sobre os Modelos Lineares Generalizados, onde se faz uma abordagem teórica mais específica, retratando as suas propriedade matemáticas mais relevantes.

O quarto capítulo é dedicado à construção da tarifa *a priori* através dos Modelos Lineares Generalizados (MLG), onde foram realizadas considerações acerca dos resultados obtidos. Inicialmente, procedeu-se a uma análise dos dados, tendo-se efectuado a depuração e a segmentação dos dados em grupos homogéneos. No final deste capítulo, apresenta-se a estrutura tarifária, onde são realizadas algumas análises e considerações sobre o modelo, avaliando o nível de adequação e qualidade de ajuste. Por fim, apresentam-se algumas, conclusões finais.

Capítulo 2

Modelos de Tarifação

2.1 Introdução

Sistemas de Tarifação são um conjunto de técnicas em que se baseia a elaboração ou construção de uma tarifa. Todo o sistema de tarifação tem o objectivo de obtenção de prémios de forma que se corresponda com o sinistro a pagar. Por isso, a tarifação constitui-se numa actividade fundamental no negócio das Seguradoras.

Entende-se por Tarifa a tabela ou quadro que contém os prémios comerciais dos distintos riscos, assim como as normas de aplicação.

Na elaboração da tarifa deve-se considerar os factores de risco mais significativos, ou seja, aqueles que melhor explicam o comportamento da sinistralidade. Estes factores devem ser segmentados em níveis tarifários consistentes para reduzir a dispersão de sinistro nas classes de risco que figurem na tarifa.

É fundamental que o cálculo do prémio seja equitativo a cada risco, tendo em conta a solvência. A equidade implica que cada segurado pague de acordo com o risco que lhe corresponde e a solvência implica assegurar que os prémios sejam significativos e suficientes, isto é, deve permitir a rentabilidade da seguradora ao longo prazo.

Os modelos de Tarifação são classificados em, ver [27]:

- Tarifação *a priori* ou Class Rating. Permite-nos atribuir um prémio ao risco

que incorpora uma apólice sem ter necessariamente a experiência de sinistralidade em concreto. É necessário apenas conhecer certas características para atribuir uma sinistralidade esperada e, com isto, o cálculo do prémio. Dado o objectivo de equidade e suficiência do prémio, busca-se a formação de grupos de riscos homogéneos determinados pela agregação de diferentes apólices (escalões tarifários), que terão sinistralidade esperada similar e, portanto, pouca dispersão em torno do valor esperado.

- Tarifação *a posteriori* ou Experience Rating. Pode-se entender, em sentido estrito, como sendo um complemento da tarifação *a priori*, como aquela que pressupõe a existência de um prémio inicial, que se vai modificando com base na experiência de sinistralidade, para dar lugar aos prémios de períodos sucessivos. Dá lugar, por exemplo, aos Sistemas de *Bonus-Malus*. Em sentido amplo pode entender-se como a actualização das tarifas mediante a incorporação de novas informações não observáveis *a priori*.

  A justificação desse sistema é que dentro de cada classe de risco existe heterogeneidade, devido à influência de certos factores de risco não considerados (conhecidos ou desconhecidos) ou agrupamento incorreto das classes. Essa heterogeneidade será recolhida na existência de sinistralidade ao longo dos anos sucessivos. Ao considerar a experiência de cada apólice, tende-se a obter um maior grau de equidade nos prémios nos anos subsequentes, incorporando informações evolutivas dos riscos através de um sistema de bonificação e penalização (Sistemas de *Bonus-Malus*) de acordo com os resultados observados.

2.1.1 Tarifa e Estrutura Tarifária

A estimativa do prémio a cobrar pela seguradora, pela aceitação de um determinado risco, requer o conhecimento das características do segurado, que servirão de indicador do risco que irão assumir.

Em seguros de massa (por exemplo, seguros automóveis), as características do risco a segurar são semelhantes a vários segurados, em vez do cálculo de um prémio individual, é mais útil a definição de uma estrutura -*a Tarifa* - que define o prémio a pagar por uma apólice, uma vez identificada as características do risco e a estrutura tarifária diz respeito à relação existente entre os prémios de uma mesma carteira, em função das variáveis que servem de base à tarifa, ver [14].

Na construção de uma tarifa, é fundamental identificar os factores tarifários (variáveis) que influenciam o risco, de modo a criar grupos homogéneos de risco, que dão origem aos *escalões (níveis)* tarifários, de forma a estabelecer o prémio adequado a pagar por uma apólice, uma vez identificado o escalão a que pertence.

A estrutura tarifária estabelece um nível base, designado por *Segurado Padrão*, que engloba determinadas características relacionadas a cada uma das variáveis tarifárias. Este nível estima o prémio correspondente e a partir dele determinam-se os coeficientes que relacionam os restantes prémios da tarifa com prémio do Segurado Padrão, procedendo *agravamentos ou descontos*, segundo a análise de risco efectuada entre o Segurado Padrão e os restantes escalões tarifários.

2.2 Tarifação *a priori*

O processo tarifário consiste nas seguintes fases, ver [27]:

1. Dados Iniciais: Depuração dos dados. Devemos ter em conta que a aplicação de um método estatístico-matemático de análise de dados a dados de pouca qualidade é uma perda de tempo, pois as conclusões não serão satisfatórias e poderá causar danos financeiros.

2. Seleção das variáveis tarifárias e determinação de classes de tarifa: Variáveis tarifárias e a escolha dos factores de risco ou características que utilizaremos para distinguir os segurados com diferentes riscos associados e como influenciam a sinistralidade. Os factores selecionados são denominados por *Variáveis Tarifárias* e a determinação de classes de tarifa consiste em selecionar as classes ou grupos de classes das variáveis tarifárias que acabam discriminando os diferentes grupos de risco na tarifa final.

   As variáveis tarifárias serão aqueles factores de risco que a seguradora utiliza para distinguir os riscos, entre os segurados, pois são aquelas que explicam uma parte importante da sinistralidade. Dependendo da variável em estudo, podemos ter variáveis tarifárias que afetam o número de sinistros, o custo do sinistro ou ambos.

   As classes de tarifa são definidas como os níveis que distinguiremos dentro de cada variável tarifária. A determinação das classes de tarifa não é simples porque, por um lado, é bom que sejam amplas para que cada classe seja tão

significativa quanto possível, isto é, incluir um grande número de observações e, por outro lado, as classes de tarifa têm que ser tão estreitas quanto possível para fornecer um melhor ajuste.

3. Obtenção dos grupos tarifários: correspondem à obtenção de grupos homogéneos de risco, considerando a frequência e custo de sinistro;

4. Tratamento especial e adequado dos grandes sinistros: corresponde a sinistros que originam um valor bastante elevado de custo de indemnizações, imputado a um pequeno número de sinistros.

5. Cálculo do prémio para cada escalão tarifário: consiste em estimar os prémios de factores de risco que ajustam a sinistralidade para cada escalão tarifário a partir da frequência de sinistro e custo de sinistro.

6. Finalmente, realiza-se a adequação da tarifa ao mercado, tendo em conta a concorrência do mercado e segmentos populacionais para qual a cobertura é direcionada.

Estas fases não são independentes, de modo que existem modelos ou métodos que servem todos ou vários ao mesmo tempo. Abaixo descrevemos as principais técnicas utilizadas no processo de tarifação a *priori*:

- Seleção das variáveis e das classes de tarifa: Análise de *clusters*, análise discriminante, técnicas de segmentação.

- Serão adoptados modelos estatísticos e actuariais adequados (Modelos Lineares Generalizados) que permitem a escolha adequada das variáveis significativas a incluir no modelo de tarifação, bem como na definição de escalões tarifários e prémios a cobrar.

Pode-se estabelecer, também, ao estimar o prémio, estruturas aditivas ou multiplicativas, de modo que o prémio correspondente a um grupo seja calculado como a soma dos efeitos de pertencer a cada uma das classes ou como o produto.

Uma análise sobre os métodos de tarifação *a priori* pode encontrar-se em [28], [14], assim como em [7]. Dedicamos os capítulos seguintes ao estudo dos Modelos Lineares Generalizados e sua aplicação à tarifação *a priori*.

2.3 Tarifação de Seguro Não-Vida

Para a tarifação é necessária a preparação de bases técnicas que compreendam, entre outros aspectos: as informações genéricas (explicação do risco segurável, os factores de risco considerados na tarifa e os sistemas de tarifação utilizados) e informações estatísticas sobre o risco (as estatísticas utilizadas indicam o tamanho da amostra, as fontes, o método de obtenção e o período referente).

Este trabalho estará baseado no estudo de um modelo tarifário *a priori* do Seguro Automóvel, considerando o conhecimento e quantificação dos factores de risco de sinistralidade sendo um requisito essencial para que o processo de tarifação seja satisfatório. Ressalta-se que tarifar correctamente este tipo de risco, garante que a carteira global da seguradora se mantenha dentro dos limites de suficiência do prémio aceitável, tanto para o segurado como para a seguradora.

2.3.1 Prémio ou preço do seguro

Suponhamos que dispomos de uma carteira ou um conjunto de apólices cujos elementos são idênticos, ou seja, observamos o mesmo risco de um produto do Ramo Não Vida. Para cada apólice, observamos os dados referentes à *sinistralidade*, variáveis aleatórias N - *Número de Sinistros* (no período de um ano) e correspondentes X_i - *Custo* do sinistro i para $i=1,...N$. O custo total dos sinistros da carteira, S, é dado pela expressão:

$$S = \sum_{i=1}^{N} X_i.$$

Se assumimos, como hipótese, a independência entre o custo por sinistro e o número de sinistros, a esperança do *custo total* $\mathbb{E}[S]$:

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] | N = n\right] = \mathbb{E}[N]\mathbb{E}[X]. \tag{2.1}$$

A expressão (2.1) denomina-se *Prémio Puro*, é designado por PP, e calcula-se através de:

$$PP = \mathbb{E}[N]\mathbb{E}[X] \tag{2.2}$$

com

- $\mathbb{E}[N]$- Frequência de Sinistralidade e
- $\mathbb{E}[X]$- Custo médio (severidade) por sinistro.

O prémio puro é a componente base do prémio (preço) do seguro, a partir do qual a seguradora estabelecerá o cálculo de prémio adequado com base num dos princípios de cálculo do prémio:

- Princípio do Valor Esperado: $P = (1 + \alpha)\mathbb{E}[S]$, onde α é o coeficiente de carga de segurança;
- Princípio da Variância: $P = \mathbb{E}[S] + \alpha Var[S]$
- Princípio do Desvio Padrão: $P = \mathbb{E}[S] + \alpha\sqrt{Var[S]}$

Com este custo a seguradora tem garantia suficiente para fazer frente aos sinistros previstos e esperados. O cálculo deve estar baseado em informações estatísticas próprias de cada seguro, portanto, é necessário partir de uma base de dados de qualidade.

2.3.2 Factores de Risco

Num processo de tarifação *a priori*, dispõe-se dos dados de sinistralidade de uma carteira composta por todos os factores de risco de cada um dos segurados. Um dos componentes chave na elaboração de uma tarifa são os factores de risco, já que sobre eles se avalia o risco em termos quantitativos.

Existem numerosos factores de risco que são tidos em conta em modelos de tarifação. Apresenta-se abaixo os mais usuais:

- Factores relacionados com o veículo: o valor, a antiguidade, a categoria, a classe, o tipo, a marca, o modelo, o tipo de combustível, a cilindrada, a potência, o peso, a relação potência/peso, etc.
- Factores relacionados com o condutor: antiguidade da carta de condução, idade, género e o resultado da experiência passada.
- Factores relacionados com a circulação: a zona de circulação (residência), a província, o uso do veículo, a quilometragem anual, etc.

Uma das fases prévias em todo processo de tarifação e elaboração preditiva é a seleção de factores de risco que são possíveis variáveis tarifárias.

É necessário conhecer e processar o máximo volume de informação em torno do risco que tem o segurado, já que a evolução de sinistralidade é permanente e factores de risco que hoje não têm uma correlação com o risco, em um futuro podem ser factores relevantes.

Capítulo 3

Modelos Lineares Generalizados

3.1 Introdução

Os modelos de regressão procuram ajustar uma função que relacione, em termos médios, a variação de uma variável dependente Y com a variação de variáveis independentes $\mathbf{X}$, sendo que a parcela de variação da variável dependente não explicada pela variação das variáveis independentes é atribuída a uma variável aleatória denominada *erro*. Dada esta função, a partir de valores das variáveis independentes, ou covariáveis, podemos estimar o valor da variável de interesse, ou variável resposta. As formas mais simples de modelos de regressão envolvem uma dependência linear entre a resposta e as covariáveis e ainda supõem que o erro tem distribuição Normal, sendo conhecidos como Modelos de Regressão Linear.

No caso dos Modelos de Regressão Linear Multivariada, a variável resposta é expressa através de uma combinação linear das covariáveis, adicionada de um erro cuja distribuição supomos ser Normal. Temos então que tal modelo requer que a variável resposta também siga uma distribuição Normal. No entanto, no âmbito da tarifação, esta não é uma situação que se verifica com frequência.

Segundo [26], o modelo de regressão linear, desenvolvido por Legendre e Gauss no início do século XIX, foi a principal técnica de modelação estatística até meados do século XX, embora vários modelos não lineares ou não normais já tenham sido desenvolvidos em face de situações que não eram adequadamente explicadas pelo modelo linear Normal, antes desse período. Podem ser citados como exemplos os

modelos de [20] e [18] , o modelo complementar log-log para ensaios de diluição, ver [9], os modelos probit [3] e logit [1] para proporções, os modelos log-lineares para dados de contagens [2], entre outros.

Os Modelos Lineares Generalizados foram introduzidos, em 1972, por [21] e nada mais são do que uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como conceptual, a abordagem de linear generalizado. Estes modelos podem ser utilizados quando a distribuição da variável de interesse é qualquer distribuição da Família Exponencial tais como a Gaussiana, Binomial, Gama, Poisson, entre outros.

3.2 Descrição do Modelo Regressão Clássico

Suponhamos que temos uma única variável aleatória Y, com função de probabilidade ou função densidade de probabilidade *f(.)*, a qual acreditamos estar associada a um conjunto de variáveis explicativas $X_1, X_2, ..., X_p$. Suponhamos também que temos uma amostra contendo n observações sendo que para cada observação temos o par $(y_i, \mathbf{x}_i)$, onde $\mathbf{x}_i$ é o vetor coluna contendo as variáveis explicativas $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})^T$. Na sua forma mais simples, temos que os modelos de regressão linear são dados por:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, ..., n \tag{3.1}$$

onde $\mathbf{X}$ é a matriz do modelo de dimensão $n \times p$, associado a um vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ de parâmetros, e $\boldsymbol{\varepsilon}$ é um vector de erros aleatório com distribuição Normal, segundo [26].

Os Modelos Lineares Generalizados são uma extensão do modelo linear clássico (ou regressão linear). A extensão é feita em duas direcções. Por um lado, a distribuição considerada não tem de ser normal, podendo ser qualquer distribuição da Família Exponencial; por outro lado, embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado e o vector de covariaveis pode ser qualquer função diferenciável, ver [26].

3.3 Características do Modelo Linear Generalizado

A estrutura de MLG pode ser descrita como um conjunto de três componentes fundamentais, ver [21], tais como, a Componente Aleatória (ou distribuição de probabilidade de ocorrência), a Componente Sistemática e a Função de Ligação.

3.3.1 Componente Aleatória

Dado o vector de covariáveis $\mathbf{x_i}$ as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à *Família Exponencial*, ou seja, a sua função densidade de probabilidade pode-se expressar da seguinte forma

$$f(y_i;\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{3.2}$$

onde:

- $a(.), b(.)$ e $c(.)$ são funções conhecidas, que variam de acordo com a distribuição da Família Exponencial;
- θ_i é o parâmetro canónico que depende da regressão, que variará de acordo com as suas características;
- ϕ é o parâmetro de dispersão.

O valor esperado da variável aleatória Y_i obtém-se pela primeira derivada de $b(\theta_i)$,

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.3}$$

e a sua variância, como um valor proporcional à dispersão, é encontrado pela segunda derivada de $b(\theta_i)$,

$$Var(Y_i) = a(\phi).b''(\theta_i). \tag{3.4}$$

3.3.2 Componente Sistemática

Esta componente do modelo, também designada como *modelo linear*, relaciona as variáveis independentes da forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} \tag{3.5}$$

e, portanto, atribui para cada observação o *preditor linear*

$$\eta_i = \sum_{j=0}^{p} x_{ij}\beta_j = \mathbf{X}_i^T \boldsymbol{\beta},\ i = 1, 2, ..., n \tag{3.6}$$

em que $\mathbf{X}_i$ a matriz das variáveis independentes e seus parâmetros pelo vector de parâmetros $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$.

3.3.3 Função de Ligação

Num modelo de regressão linear a variável resposta pode flutuar em $\mathbb{R}$ e espera-se que seja Normalmente distribuída. O mesmo não sucede nos modelos a ser analisados neste trabalho, onde se espera que a variável independente tome valores em $\mathbb{R}^+$. Seja $E(Y_i) = \mu_i$, onde se estabelece uma transformação não-linear da média μ_i da variável Y_i tal que

$$g(\mu_i) = \eta_i,\ i = 1, 2, ..., n. \tag{3.7}$$

A função $g(.)$ é uma função monótona e diferenciável, designada por função de ligação, estabelecendo uma relação não-linear entre a variável resposta e as variáveis explicativas e a sua inversa é dada pela expressão:

$$\mu_i = g^{-1}(\eta_i),\ i = 1, 2, ..., n. \tag{3.8}$$

3.3.3.1 Modelo Aditivo e Multiplicativo

Para a construção da estrutura tarifária, é necessário definir o tipo de função de ligação a adoptar em função de se pretender optar por uma estrutura modelo Multiplicativo ou Aditivo. Podem definir-se outros modelos, mas estes são mais utilizados, ver [14] e [22].

Os modelos Multiplicativo ou Aditivo estabelecem uma relação entre o prémio do segurado padrão e os restantes segurados, da seguinte forma:

- Modelo Aditivo: A relação existente entre os prémios é estabelecida através da soma de uma quantidade monetária que reflete as diferenças de risco dos segurados, ou seja, a mudança no valor de um factor tarifário traduz-se numa alteração em valor absoluto no valor do prémio a cobrar, para todos os restantes factores tarifários.

- Modelo Multiplicativo: a relação existente entre o prémio é estabelecida através de uma proporção do prémio do segurado padrão, em função das diferenças

de risco entre os escalões (níveis), ou seja, a mudança no valor de um factor tarifário traduz-se numa alteração proporcional no valor do prémio a cobrar, independentemente dos restantes factores tarifários.

Para o presente trabalho, adoptar-se-á o modelo multiplicativo, que corresponde à utilização da função logarítmica para a função de ligação.

3.3.4 Parâmetro de Exposição ao Risco

A inclusão de um parâmetro de *Exposição ao Risco*, num modelo de tarifação, é um procedimento comum e fundamental, uma vez que o tempo de exposição ao risco para cada apólice é diferente. Este parâmetro é importante na modelação do Número de Sinistros, em que o tempo de exposição ao risco em carteira se revele de extrema importância nas estimativas da frequência de sinistralidade. Na área de seguros, o período de registo dos dados coincide com o ano civil, enquanto os períodos de vigência das apólices anuais estão desfasados com este. Assim, as informações registadas refletirão apenas uma parte da validade total do contrato. Além de poder introduzir a exposição ao risco como um valor de ponderação w_i, também pode ser incorporado como um efeito conhecido ao preditor linear η_i, mediante a inclusão de ξ_i, de seguinte maneira:

$$\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} + \xi_i,\ i = 1, 2, ..., n \tag{3.9}$$

de onde se estabelece a relação com o valor esperado da variável aleatória Y_i

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta} + \xi_i). \tag{3.10}$$

3.4 Estimação do Modelo

A estimação dos parâmetros do modelo requer a realização de inferência sobre os dados. Em [21] propõe-se, em primeiro lugar, determinar uma medida de adequação entre os valores estimados pelo modelo e os dados. Os valores de máxima verosimilhança dos parâmetros do modelo serão aqueles que minimizam o desvio, que será analisado mais adiante. Para alcançar este objetivo, em segundo lugar, propõe-se utilizar um algoritmo iterativo de mínimos quadrados ponderados e como, em geral, não há equações que satisfaçam essas características ótimas, será necessário realizar a estimação por meio de métodos numéricos.

3.5 Tipo de Modelos

O conhecimento prévio da distribuição para a variável resposta (tal como referido na secção 3.3.1) é fundamental quando se pretende efectuar a modelação através dos Modelos Lineares Generalizados. No que se segue, apresentam-se alguns resultados das distribuições mais usuais no âmbito da tarifação.

3.5.1 Modelo de Poisson

A distribuição de Poisson, por apresentar probabilidade de sucesso reduzido e ser adequada à modelação de dados de contagem, é frequentemente utilizada para modelar o Número de Sinistros automóveis. As propriedade subjacentes a uma distribuição de Poisson, segundo [16], são:

- As observações estudadas são homogéneas;
- A ocorrência de um sinistro é um evento raro, a probabilidade de ocorrência é muito pequena;
- Não existe contágio, ou seja, a ocorrência de um sinistro posterior não é influenciado pelo anterior.

Seja Y uma variável aleatória com distribuição de Poisson, isto é, $Y \sim P(\mu)$, e com função de probabilidade dada pela expressão:

$$f_{Y_i}(y_i, \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = exp\left[\frac{y_i.ln(\mu_i) - e^{ln(\mu_i)}}{1} + (-ln(y_i!))\right] \tag{3.11}$$

pelo que se pode afirmar que a distribuição de Poisson pertence à Família Exponencial, com $b(\theta_i) = e^{\theta_i}$, $a(\phi) = 1$, $c(y_i, \theta) = -ln(y_i!)$.
De acordo com (3.3), obtém-se a média da distribuição pela expressão

$$E(Y_i) = b'(\theta_i) = (e^{\theta_i})' = e^{\theta_i} = e^{ln(\mu_i)} = \mu_i \ .$$

Relativamente à variância, temos de acordo com (3.4),

$$Var(Y_i) = a(\phi).b''(\theta_i) = 1.(e^{\theta_i})'' = e^{\theta_i} = e^{ln(\mu_i)} = \mu_i$$

A estimação do modelo faz-se mediante o método de máxima verosimilhança. A função de verosimilhança, obtida a partir da distribuição de Poisson, será:

$$\mathcal{L}(y, \mu) = \prod_{i=1}^{n} f_{Y_i}(y_i; \mu_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}. \tag{3.12}$$

Para facilitar as operações utiliza-se a função de log-verosimilhança, em ambos os membros da equação (3.12), obtendo-se:

$$log\mathcal{L}(y,\mu) = \sum_{i=1}^{n}[y_i ln(\mu_i) + \mu_i - ln(y_i!)]. \tag{3.13}$$

Introduzindo a relação com as covariáveis, mediante a função de ligação canónica, que no caso Poisson é a função logaritmo, tem-se $\mu_i = exp\left(\sum_{j=1}^{p} x_{ij}\beta_j + \xi_i\right)$, e obtém-se:

$$log\mathcal{L}(y,\mu) = \sum_{i=1}^{n}\left[y_i.\sum_{j=1}^{p}(x_{ij}\beta_j + \xi_i) - \left[\sum_{j=1}^{p}(x_{ij}\beta_j + \xi_i)\right] - ln(y_i!)\right]. \tag{3.14}$$

O valor da log-verosimilhança é máximo quando para cada valor j, a derivada parcial de primeira ordem da função de log-verosimilhança em relação a β_j for igual a 0

$$\frac{\partial log\mathcal{L}}{\partial \beta_j} = 0, j = 1, 2, ..., p$$

A obtenção da estimativa de máxima verosimilhança requer a resolução de um sistema de equações não lineares. Devido ao grande número de observações utilizadas, recorre-se a método numéricos iterativos, para a obtenção das estimações.

A não verificação de algumas das principais propriedades da distribuição de Poisson, nas carteiras de seguros, produzirá diferentes anomalias:

- Sobre-dispersão: O facto da variância ser superior à média é chamado de *sobre-dispersão*, uma vez que na sinistralidade gerada por uma distribuição de Poisson, a média deveria ser igual à variância. Este efeito é geralmente comum em dados de carteiras de seguros.
- Inflação de zeros: Essa anomalia é observada quando os dados da amostra apresentam uma frequência mais elevada para a ocorrência de zero sinistros do que seria esperado se a amostra tivesse sido gerada por uma distribuição de Poisson.

3.5.2 Modelo de Binomial Negativa

O problema da sobre-dispersão, referido no modelo de Poisson, leva a buscar distribuições que admitam este tipo de comportamento, como é o caso da Binomial Negativa (BN), frequentemente utilizada para modelar o número de sinistros automóveis, quanto a variância da carteira é superior à média.

Seja Y a v.a. que contabiliza o número de provas de Bernoulli a realizar, até obter-se r sucessos. Diz-se que Y segue uma distribuição Binomial Negativa, $Y \sim BN(r,p)$, sendo a função de probabilidade dada por

$$f(y) = \begin{pmatrix} r+y-1 \\ r-1 \end{pmatrix} p^r (1-p)^y. \tag{3.15}$$

com p a probabilidade de sucesso.
A função de probabilidade da distribuição Binomial Negativa pode ser generalizada para $r > 0$, fazendo:

$$f(y) = \frac{\Gamma(y+r)}{y!\,\Gamma(r)} p^r (1-p)^y. \tag{3.16}$$

É usual utilizar-se uma reparametrização da distribuição fazendo $\mu = \dfrac{r(1-p)}{p}$ e $k = \frac{1}{r}$. Desta feita, a função de probabilidade será:

$$f(y) = \frac{\Gamma(y+\frac{1}{k})}{y!\,\Gamma(\frac{1}{k})} \left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}} \left(\frac{k\mu}{1+k\mu}\right)^{y}. \tag{3.17}$$

Assim, a média e a variância são, respectivamente,

$$E(Y) = \mu \tag{3.18}$$

$$Var(Y) = \mu(1+k\mu) \tag{3.19}$$

Para a estimação dos parâmetros da Binomial Negativa, recorre-se ao método da máxima verosimilhança, em que a função de verosimilhança é dada pela expressão:

$$\mathcal{L}(y,\mu_i) = \prod_{i=1}^{n} \left[\frac{\Gamma(y+\frac{1}{k})}{y_i!\,\Gamma(\frac{1}{k})} \left(\frac{1}{1+k\mu_i}\right)^{\frac{1}{k}} \left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i} \right]. \tag{3.20}$$

Aplicando a função log-verosimilhança, e tendo em conta que $\mu_i = e^{\mathbf{X}\boldsymbol{\beta}}$, obtém-se a expressão:

$$log\mathcal{L}(y, e^{\mathbf{X}\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left[ln\Gamma\left(y_i + \frac{1}{k}\right) - ln(y_i!\,) - ln\Gamma\left(\frac{1}{k}\right) - \left(y_i + \frac{1}{k}\right) ln(1+ke^{\mathbf{X}\boldsymbol{\beta}}) + y_i ln(ke^{\mathbf{X}\boldsymbol{\beta}}) \right].$$

O valor da função log-verosimilhança é máximo quando para cada valor, as derivadas parciais de primeira ordem da função de log-verosimilhança em ordem a β_j e k forem iguais a 0, isto é

$$\frac{\partial log\mathcal{L}}{\partial \beta_j} = 0$$

e

$$\frac{\partial log\mathcal{L}}{\partial k} = 0$$

A obtenção das estimativas de máxima verosimilhança $\hat{\beta}$ e $\hat{k}$ explicitamente requer resolver um sistema de equações não lineares, em se obtém soluções aproximadas mediante métodos numéricos.

3.5.3 Modelo Gama

Admitindo que as v.a´s $Y_i \sim Gama(\alpha, \beta)$ são independentes e um modelo dos MLG adequado para variáveis respostas contínuas, no caso em que esta é estritamente positiva e apresenta assimetria à direita, como é usualmente o caso dos montantes a pagar por sinistros decorrentes de acidentes automóveis. A função de densidade de probabilidade de distribuição $Gama(\alpha, \beta)$ é dada por

$$f(y, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0. \tag{3.21}$$

Parametrizando em $\mu = \alpha/\beta \Leftrightarrow \beta = \alpha/\mu$, a função de probabilidade pode escrever-se da seguinte forma:

$$\begin{aligned} f(y, \alpha, \beta) &= \frac{(\alpha/\mu)^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-(\frac{\alpha}{\mu})y} = \\ &= exp\left\{ \frac{\frac{-1}{\mu} y - log(\mu)}{1/\alpha} + (\alpha - 1)log(y) - log(\Gamma(\alpha)) \right\} . \end{aligned}$$

Com esta expressão prova-se que a distribuição Gama pertence a Família Exponencial, tendo a média e a variância, respectivamente,

$$\mathbb{E}(Y) = \frac{\alpha}{\beta} \tag{3.22}$$

$$Var(Y) = \frac{\alpha}{\beta^2} \tag{3.23}$$

A obtenção da estimativa é feita através do método da máxima verosimilhança, conforme procedido nas secções 3.5.1 e 3.5.2.

Na Tabela 3.1 encontram-se algumas distribuições de probabilidade da Família Exponencial, frequentemente utilizadas na modelação de tarifação, ver [14].

Tabela 3.1: Algumas Distribuições da Família Exponencial

Distribuições		θ	$b(\theta)$	ϕ	$\mathbb{E}[Y]$	$Var[Y]$
Normal	N(μ,σ^2)	μ	$\frac{\theta^2}{2}$	σ^2	μ	1
Binomial	B(n,π)	$\ln(\frac{\pi}{1-\pi})$	$n\ln(1+e^{\theta})$	1		$n\pi(1-n\pi)$
Poisson	P(μ)	$\ln(\mu)$	e^{θ}	1	μ	μ
Binomial Negativa	B(μ,k)	$\ln(\frac{k\mu}{1+\mu})$	$-\frac{1}{k}ln(1-ke^{\theta})$	1	μ	$\mu(1-k\mu)$
Gama	G(α,β)	$-\frac{1}{\mu}$	-ln($\frac{\alpha}{\mu}$)	$\frac{1}{\alpha}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

3.5.4 Modelo Lognormal

Uma variável aleatória Y tem distribuição Lognormal com parâmetros μ e σ^2, escrevendo-se $Y \sim LN(\mu, \sigma^2)$ quando $X = Log(Y)$ segue uma distribuição Normal. Logo, a sua função densidade de probabilidade é dada por

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} exp\left\{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right\}, y \in \mathbb{R}_0^+, \mu, \sigma \in \mathbb{R} \tag{3.24}$$

A média e a variância são, respectivamente,

$$\mathbb{E}(Y) = exp\left\{\mu + \frac{\sigma^2}{2}\right\} \tag{3.25}$$

$$Var(Y) = exp\left\{2\mu + \sigma^2\right\}\left[exp(\sigma^2) - 1\right]. \tag{3.26}$$

A distribuição Lognormal não pertence à Família Exponencial, que consiste no ajustamento de uma regressão linear clássica multivariada ao logaritmo da variável aleatória, usado para modelar variáveis não negativas com assimetria positiva, como por exemplo, do Custo de um Sinistro.

3.5.5 Distribuição Pareto Generalizada

A distribuição Pareto Generalizada (GPD) destaca-se como sendo uma distribuição utilizada na modelação de observações extremas. A GPD foi introduzida por Pickands (1975) como uma distribuição para modelação dos excessos de uma amostra. Com aplicação em análise de eventos extremos, na modelação de grandes sinistros de seguros e em qualquer situação que a distribuição exponencial pode ser utilizada, mas que é necessário alguma robustez contra alternativa de cauda mais pesada ou mais leve, segundo [25] e [15].

Considera-se uma variável aleatória Y segue um distribuição GPD, com parâmetros de forma, localização e escala designados por ε, $\mu \in \mathbf{R}$ e $\sigma > 0$, respectivamente, se a sua função densidade de probabilidade é dada pela expressão

$$f(y|\varepsilon,\mu,\sigma) = \begin{cases} \frac{1}{\sigma}\left(1+\varepsilon\frac{y-\mu}{\sigma}\right)^{-\frac{1}{\varepsilon}-1}, & \varepsilon \neq 0 \\ \frac{1}{\sigma}exp\left(-\frac{y-\mu}{\sigma}\right), & \varepsilon = 0 \end{cases}$$

Se $\varepsilon > 0$ então $y \geq \mu$, enquanto que, se $\varepsilon < 0$ temos $\mu \geq y \geq \mu - \frac{\sigma}{\varepsilon}$.

A distribuição Pareto Generalizada é uma distribuição bastante versátil, dado que muitas distribuições de probabilidade são casos particulares desta, consoante o valor que os seus parâmetros tomam, por exemplo:

- Se $\varepsilon = 0$ e $\mu = 0$, reduz-se à distribuição Exponencial de valor médio σ.
- Se $\varepsilon = 0$ e $\mu = \sigma/\varepsilon$, coincide com distribuição Pareto com parâmetro de forma $1/\varepsilon$ e escala σ/ε.
- Se $\varepsilon = -1$ e $\mu = 0$, reduz-se á distribuição Uniforme em (0,σ).

Segundo [8], a média e a variância são dados pela expressão:

$$\mathbb{E}(Y) = \mu + \frac{\sigma}{1-\varepsilon},\ \varepsilon < 1\ . \tag{3.27}$$

$$Var(Y) = \frac{\sigma^2}{(1-\varepsilon)^2(1-2\varepsilon)},\ \varepsilon < 1/2\ . \tag{3.28}$$

3.6 Qualidade de Ajustamento

A qualidade de ajustamento consiste em justificar em que medida o modelo estimado se ajusta ao conjunto de observações. Essas medidas podem ser usadas para estimar os parâmetros do próprio modelo, de modo a comparar a precisão de diferentes modelos.

3.6.1 Erro Quadrático Médio

O Erro Quadrático Médio (EQM), é uma medida de qualidade de ajuste que mostra o desvio da variável estimada sobre a variável observada como média dos erros ao

quadrado

$$EQM = \frac{\sum_{i=1}^{n}(Y - \hat{Y})^2}{n},$$

pelo que, um menor valor indica um melhor ajuste do modelo aos dados.

3.6.2 Deviance

A Deviance (ou Desvio) é uma estatística que mede a qualidade do modelo, em que um valor mais alto indica um pior ajuste. É uma generalização da Soma dos Quadrados dos Resíduos da estimação pelo Métodos dos Mínimos Quadrados Ordinários.

Esta medida parte do princípio que para observar a eficácia do modelo se deve comparar com um modelo mais amplo, que contenha o máximo número de parâmetros sem repetições possíveis, chamado de modelo saturado (M_s). Este modelo M_s será muito similar ao modelo que se trata de medir M_p (modelo proposto ou estudado), já que conterá a mesma distribuição e a mesma função de ligação. O valor da função de verosimilhança $\mathcal{L}M_s$ será maior do que qualquer outra função para estes dados, pois fornecerá a descrição mais completa e precisa dos dados. Ao compará-lo com o valor da função de verosimilhança do modelo $\mathcal{L}M_p$ obter-se-á o seguinte rácio de verosimilhança:

$$\lambda = \frac{\mathcal{L}(M_s)}{\mathcal{L}(M_p)} \, . \tag{3.29}$$

Na prática, [21] propõe que se utilize o logaritmo na expressão anterior, obtém-se

$$ln(\lambda) = \mathcal{L}(M_s) - \mathcal{L}(M_p) \, .$$

O desvio é, portanto, definido como uma medida de afastamento entre o modelo saturado e o modelo proposto, calculado como

$$D = 2\left[\mathcal{L}(M_s) - \mathcal{L}(M_p)\right] \, . \tag{3.30}$$

3.6.3 Critério de Informação de Akaike

O Critério de Informação de Akaike (AIC), é outra medida de qualidade de ajuste do modelo de grande utilidade na comparação de modelos distintos, uma vez que não considera apenas o ajuste do modelo aos dados, mas também introduz uma componente de penalização sobre o aumento do número de parâmetros do modelo. Essa penalização tenta evitar sobre-ajuste, o que pode ocorrer ao introduzir um excesso de parâmetros no modelo.

Seja $\mathcal{L}M_p$ o máximo da função de verosimilhança do modelo proposto e p o número de parâmetros do mesmo modelo. A estatística de AIC será:

$$AIC = -2\ ln(\mathcal{L}(M_p) + 2p)$$

pelo que um valor menor de AIC indicará, em princípio, que se trata de um modelo de maior qualidade que o outro com um AIC maior.

3.6.4 Análise de Resíduos

A análise dos resíduos consiste em técnicas destinadas a verificar a validade das hipóteses efectuadas sobre o modelo, relativamente à escolha da distribuição, da função de ligação e em termos do preditor linear, como também ajuda a identificar observações mal ajustadas, que não são explicadas pelo modelo, ver [14].

Um resíduo R_i deve exprimir a discrepância entre o valor observado y_i e o valor $\hat{\mu_i}$ ajustado do modelo. No modelo linear normal, em que o vector das respostas $\mathbf{Y}$ se podem escrever como

$$\mathbf{Y} = \boldsymbol{\mu} + \varepsilon = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

tem-se $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ e o vector dos resíduos é naturalmente dado por $\boldsymbol{R = y - \hat{y}}$ onde $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ é o vector dos valores ajustados, ver [26].

A escolha mais comum para avaliar as observações são os *Resíduos de Pearson*, segundo [22], dados pela expressão

$$R_{P_i} = \frac{y_i - \hat{\mu_i}}{\sqrt{var[\hat{\mu_i}]/w_i}},\ i = 1, ...n\ . \tag{3.31}$$

em que w_i é a exposição ao risco.
A soma destes resíduos coincide com a estatística do χ^2 de Pearson não reduzida, isto é, $\sum_{i=1}^{n} R_{P_i}^2 = \phi\chi^2$, ver [22]. Se os valores médios μ_i fossem conhecidos, os resíduos de Pearson teriam valor esperado nulo e variância constante igual a ϕ.

Os *Resíduos Standartizados de Pearson* são resíduos de Pearson divididos por $\sqrt{\phi(1 - h_i)}$ sendo, portanto, definidos por

$$R_{P_i}^* = \frac{R_{P_i}}{\sqrt{1 - h_i}} \tag{3.32}$$

onde h_i é a correção pelo facto que μ_i é estimado, sendo h_i o elemento da diagonal principal da matriz

$$\mathbf{D}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^{1/2}$$

onde $\mathbf{D} = diag\left(\frac{w_i}{\phi}var[\mu_i]\right)$.

Alternativamente, pode-se usar os resíduos em função de desvios, definidos por

$$R_{D_i} = \sqrt{w_i d(y_i, \hat{\mu}_i)} * sinal(y_i - \hat{\mu}_i) \tag{3.33}$$

onde $sinal(y_i - \hat{\mu}_i)$ é a função que indica que se deve adoptar o sinal (positivo ou negativo) do seu argumento e $d(y_i, \hat{\mu}_i)$ corresponde ao desvio de cada observação.

Os resíduos são analisados graficamente, através de gráfico de índice (em relação ao número de observações), com intuito de encontrar valores discrepantes que podem ser erróneos ou não adequados para o modelo, ver [22].

3.7 Selecção do Modelo

O objectivo deste processo é incluir no modelo um número mínimo de variáveis necessárias para realizar uma melhor explicação dos dados o que apresenta, entre outros, os benefícios: previne a sobre-ajuste, reduz a dificuldade de estimação do modelo, etc. Um dos métodos mais comum é o método de *Stepwise*, que pode ter duas dinâmicas diferente, *Forward Stepwise* (seleção progressiva) e *Backward Stepwise* (seleção regressiva), para mais detalhes ver [14] e [22].

3.7.1 Testes de Hipóteses

Após a estimação dos coeficientes de regressão é fundamental avaliar a significância das covariáveis que deverão integrar o modelo, isto é, determinar se as variáveis independentes introduzidas no modelo estão significativamente associadas à variável resposta. Para tal, recorremos ao Teste de Wald e ao Teste de Razão Verosimilhanças, ver [26] e [22].

Teste de Wald

O teste de Wald é utilizado quando se pretende testar a hipótese nula de significância ou de nulidade estatística para um subconjunto particular de coeficientes do vector de parâmetro β_j estimados. As hipóteses a testar são:

$$H_0 : \beta_j = 0 \; versus \; H_1 : \beta_j \neq 0, \; j = 1, ..., p$$

A estatística de teste e a respectiva distribuição, sob a validade de H_0, são:

$$W = \frac{\hat{\beta}_j^2}{\sigma_{jj}} \sim \chi_1^2.$$

Rejeita-se a hipótese nula, a um nível de significância α, se o valor observado da estatística for superior ao quantil de probabilidade $1-\alpha$ da distribuição χ_1^2.

Teste de Razão Verosimilhança

O Teste de Razão de Verosimilhanças é utilizado para comparar a qualidade do ajustamento de dois modelos encaixados, isto é, modelos em que um tem um subconjunto de variáveis do outro modelo. Também se pode dizer que este teste avalia a significância dos coeficientes estimados simultaneamente, ou seja, verifica se o modelo estimado é globalmente significativo.

A estatística de teste e a respectiva distribuição, sob a validade de H_0, ver [26], é definida por:

$$\Lambda = -2ln\left[\frac{max_{H_0}L(\beta)}{max_{H_0\cup H_1}L(\beta)}\right] = -2\left[\ell(\tilde{\beta}) - \ell(\hat{\beta})\right] \sim \chi_r^2. \tag{3.34}$$

onde $\tilde{\beta}$, o estimador de máxima verosimilhança restrito, é o valor de β que maximiza a verosimilhança sujeito às restrições impostas pela hipótese $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$.

As técnicas apresentadas neste capítulo serão aplicados, ao longo do próximo capítulo, na construção da estrutura tarifária que melhor representa o risco da carteira de Responsabilidade Civil Automóvel em estudo.

Capítulo 4

Construção do Modelo Tarifário

Neste capítulo, estudar-se-ão as características de uma carteira de Responsabilidade Civil Automóvel de uma Seguradora Angolana e propor-se-á uma tarifa adequada ao risco observado, mediante a aplicação de Modelos Lineares Generalizados, mediante a utilização do software R Project.

Os dados utilizados neste trabalho correspondem a 60 000 apólices do Seguro de Responsabilidade Civil Automóvel, observados no espaço temporal de 4 anos, e facultados por uma seguradora Angolana. A base de dados fornecida comporta 12 variáveis, como se ilustra na Tabela 4.1.

Tabela 4.1: Variáveis para a Tarifação

	Variáveis	Descrição
Dados da Apólice	IDap	Identificador da Apólice
	ExpRisco	Exposição ao risco
Dados do Veículo	cilindrada	Cilindrada do Veículo
	lugares	Número de Lugares do Veículo
	tipoVeic	Tipo de Veículo
	tipoUso	Tipo de Uso do Veículo
	zonaCir	Zona Habitual de Circulação do Veículo
	idadeVeic	Idade do Veículo
Dados do Segurado	AnoCart	Ano da Carta
	idadeCond	Idade do Condutor
Dados de Sinistralidade	sinistro	Nº de Sinistros numa Anuidade
	custo	Custo do Sinistro

4.1 Identificação

Dado o conjunto de variáveis, a análise focaliza-se em efetuar a inferência do número de sinistros a partir das observações da variável "sinistro" e a sua relação com as restantes variáveis, ou seja, o modelo de frequência de sinistralidade terá como variável resposta "Número de Sinistro" (variável quantitativa discreta) e o modelo de severidade ou custo por sinistro tem como variável resposta "custo" (variável quantitativa continua).

A variável "ExpRisco" (variável contínua) irá funcionar como um termo *offset*, correspondente ao período de vigência da apólice no período de observação. Esta é uma variável indispensável para o estudo de perfil de risco de um segurado, pois o período de exposição ao risco pode ter um impacto significativo na estimativa da frequência de sinistralidade.

Analisando as variáveis principais deste trabalho, que assenta uma base de dados com 60 000 apólices de Responsabilidade Civil Automóvel, observa-se que 2,75% (1 647 apólices) participaram sinistros, como ilustra a Tabela 4.2.

Tabela 4.2: Nº de Apólices e Total de Sinistros

Nº Apólice	Nº de Sinistros	Taxa de sinistralidade
60 000	1 647	2,75%

4.2 Depuração e Transformação de Dados

O objectivo deste passo consiste em apurar e efectuar o tratamento dos dados, de modo a dispor de uma versão viável da base de dados, sem erros e observações atípicas. Apresentamos na Tabela 4.3, a distribuição do Número de Sinistros por ano, considerando os dados entre 2012 e 2015.

Tabela 4.3: Número de Sinistros por Ano (2012-2015)

Nº sinistros	0	1	2	3	5
Nº Apólices	58 353	1 568	73	4	2
Proporção (%)	97,255	2,613	0,122	0,007	0,003

Nesta averiguação, determinou-se os valores máximos e mínimos de cada uma das covariáveis (como ilustrado na Tabela 4.4), para verificar se existem valores atípicos.

Tabela 4.4: Máximo e Mínimo das Covariáveis

	Nº de lugares do Veículo	Idade do Veículo	Anos de Carta	Idade do Condutor	Número de Sinistros	Custo por Sinistros ($)
Mín	0	3	4	23	0	0
Máx	950	58	69	99	5	101 504

Durante o processo de depuração, verificou-se a existência de apólices que apresentavam valores anormais ou incoerentes, que foram removidos da base de dados. Esta remoção provocou a redução de número de apólices para 59 466 (uma perda na ordem dos 0.89%), sendo que apenas 1 629 declararam sinistros, que corresponde a 2,73% da taxa sinistralidade, como ilustrado nas Tabela 4.5 e 4.6.

Tabela 4.5: Frequência de Número de Sinistro

Nº sinistros	0	1	2	3	5	**Total**
Nº Apólices	57 837	1 551	72	4	2	**59 466**

Tabela 4.6: Nº de Apólices e Total de Sinistros após Depuração

Nº Apólice	Nº de Sinistros	Taxa de sinistralidade
59 466	1 629	2,73%

4.3 Análise Descritiva e Segmentação das Variáveis

Nesta secção apresentamos o estudo dos dados mediante uma análise estatística descritiva, que consiste em efectuar uma análise descritiva univariada que tem por objetivo descrever, caracterizar e extrair conclusões individuais de cada uma das variáveis, bem como uma representação gráfica das mesmas.

4.3.1 Número de Sinistros

O Número de Sinistros é a variável resposta para o modelo de Frequência de Sinistralidade e é sobre ela que se pretende analisar o efeito das covariáveis. Trata-se de uma variável quantitativa discreta composta por números não negativos cujas

estatísticas principais se encontram na Tabela 4.7.

Tabela 4.7: Estatísticas Descritivas do Número de Sinistros por Apólice

Nº Apólices	Média	Variância	Coef. Assimetria
59 466	0,02887	0,03154	6,9692

Observando a Tabela 4.7, verifica-se que o número médio de sinistros é de 0,0289, ou seja, o número médio de sinistros que um segurado participou durante uma anuidade (a média é aqui entendida, como um valor global da carteira para um horizonte de um ano). Analisando a relação entre a média e a variância, verifica-se que a variância é maior que a média, o que demonstra sobre-dispersão da distribuição da variável aleatória. O coeficiente de assimetria com valor positivo indica que existe uma assimetria positiva com respeito à média.

O gráfico da Figura 4.1 apresenta uma melhor visualização da distribuição de número de apólices por número de sinistros. Por se tratar da variável resposta para a

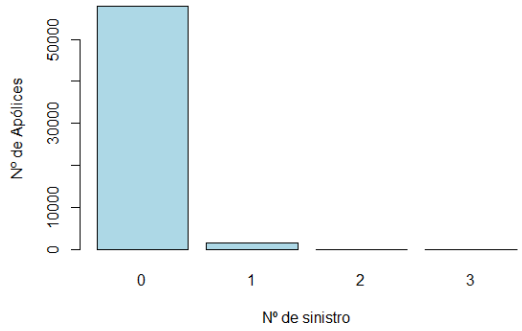
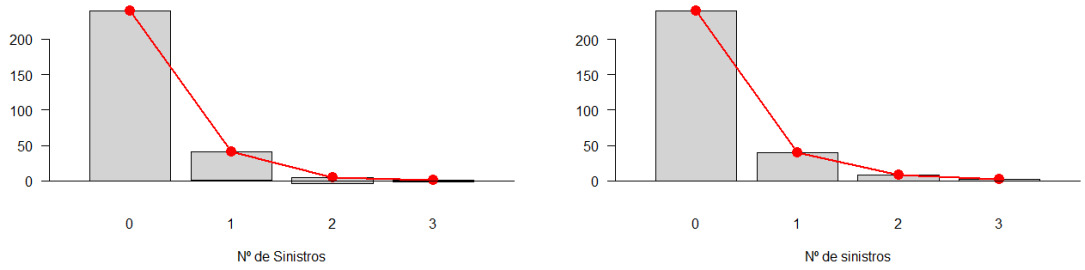


Figura 4.1: Nº de Sinistros por Apólice

modelação de frequência de sinistralidade, efectuaremos o teste de ajustamento aos dados através do Teste de Qui-Quadrado, para aferir a distribuição que melhor se ajusta aos dados, considerando-se a distribuição Poisson e Binomial Negativa, como possíveis distribuições adequadas a este fenómeno, como confirma [17] ou [19].

O teste de ajustamento de Qui-Quadrado, permite testar a hipótese de que uma determinada amostra aleatória tenha sido extraída de uma distribuição específica, a um determinado nível de significância. Portanto, para determinar a distribuição adequada ao Número de Sinistros, aplicou-se o referido teste, com parâmetros estimados pelo método da máxima verosimilhança, obtendo $p - value$ de 5.603206e-22

para modelo Poisson, o que leva a rejeitar a hipótese de que os dados tenha sido proveniente da distribuição de Poisson. Para o modelo da distribuição Binomial Negativa, obteve-se $p - value$ de 0,2432702, significando que não se rejeita a hipótese de que os dados seguem uma distribuição Binomial Negativa, aos níveis de significância usuais. Esta tese é corroborada pelos gráficos da Figura 4.2, em que se verifica um melhor ajustamento dos dados à distribuição Binomial Negativa do que à distribuição Poisson.



(a) Modelo Poisson (b) Modelo Binomial Negativa

Figura 4.2: Ajustamento dos Modelos

4.3.2 Custo por Sinistro

A Variável Custo por sinistro é a variável resposta na modelação de custo médio por sinistro e é sobre ela que se pretende avaliar o efeito das covariáveis. Das 1 629 apólices que declararam sinistros, 1 380 foram indemnizadas, sendo o custo máximo observado na carteira de 101 504 $, como ilustra a Tabela 4.8. Na Tabela 4.9, ilustram-se como os custos associados a sinistros estão distribuídos ao nível dos quantis e verifica-se que 99% dos sinistros custaram no máximo 37.446,25 $.

Tabela 4.8: Estatística de Custo por Sinistro

Nº Apólices	Nº Apólice com custo	Custo médio	Desvio Padrão	Valor ($)	
				Mínimo	Máximo
59 466	1 380	4 093,5 $	7 758,9 $	27	101.504

Tabela 4.9: Quantis dos Custos Totais ($)

50%	90%	95%	99%
1 811	9 470,2	15 837,3	37 446,3

Havendo na carteira sinistros que originaram valores bastante elevados de custos com indemnização, o que poderá influenciar a modelação de custo médio por sinistro, procedeu-se à separação dos sinistros em dois tipos: *Sinistros Regulares* e *Grandes Sinistros*. Tendo em conta o custo limite de 15 000 $, acima do qual se encontram os "Grandes" sinistros e abaixo os sinistros "Regulares", ou seja,

- Sinistros Regulares: Montante de sinistro igual ou inferior a 15 000 $.
- Grandes Sinistros: Montante de sinistro superior a 15 000 $, que será estudado no final do processo de tarifação.

Assim, para s=15.000 $, a estimativa do custo médio por sinistro é dada pela expressão:

$$\mathbb{E}[Y|\mathbf{X}] = \underbrace{\mathbb{E}[Y|\mathbf{X}, Y \leqslant s]}_{A}\underbrace{\mathbb{P}[Y \leqslant s|\mathbf{X}]}_{1-C} + \underbrace{\mathbb{E}[Y|\mathbf{X}, Y > s]}_{B}\underbrace{\mathbb{P}[Y > s|\mathbf{X}]}_{C} \tag{4.1}$$

em que:

- A: Custo médio de sinistros "Regulares".
- B: Custo médio de um "Grande" sinistro.
- C: Probabilidade de ocorrência de um "Grande" sinistro.

O valor esperado do custo médio por sinistros "Regulares", será estimado através dos Modelos Lineares Generalizados, enquanto que a probabilidade de ocorrência de um "Grande" sinistro, será estimada por meio de um Modelo Binomial (Regressão Logística) ou considera-se a probabilidade de ocorrência igual para todos os segurados, caso o modelo de regressão logística apresente factores tarifários não significativos.

Relativamente, ao valor esperado de custo médio de um "Grande" sinistro, procede-se-á à modelação da variável aleatória, através do Modelo Linear Generalizado, caso a distribuição adequada à severidade deste sinistro pertença à Família Exponencial. Caso contrário , calcula-se o valor esperado da distribuição que melhor se adequa aos dados.

4.3.2.1 Custo com Sinistros Regulares

Com base na separação efectuada, verifica-se que das 1 380 apólices indemnizadas, 1 307 apresentam custo de indemnização igual ou inferior a 15 000 $, com custo médio por sinistro de 2 696,05 $ e o custo máximo de 14 985 $, conforme mostra a Tabela 4.10. Segundo quantis de custos totais, constata-se que 99% dos sinistros custaram no máximo 14 133,3 $, de acordo com a Tabela 4.11.

Tabela 4.10: Estatística de Custo por Sinistros Regulares

Nº Apólices	Nº Apólice com custo	Custo médio	Desvio Padrão	Valor ($)	
				Mínimo	Máximo
59 466	1 380	2 696,05 $	2 943,86 $	27	14 985

Tabela 4.11: Quantis dos Custos por Sinistros Regulares ($)

50%	90%	95%	99%
1 653	6 607,20	9 436,20	14 133,28

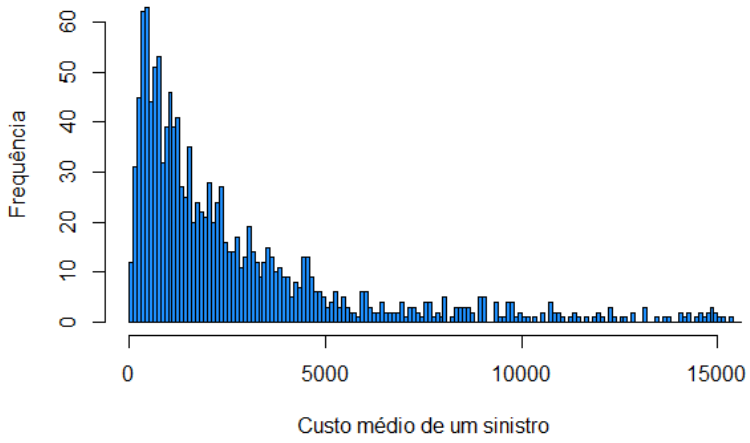


Figura 4.3: Histograma do Custo por Sinistro

Verificando o histograma da Figura 4.3, observa-se uma assimetria positiva, o que demonstra a existência de muitos sinistros com custo baixo e poucos com custo alto, que nos auxilia na escolha de uma distribuição para variável resposta. Para o efeito, utilizamos o teste de ajustamento de Kolmogorov-Smirnov, que testa a hipótese se uma amostra é proveniente de uma determinada distribuição, baseada na comparação da função de distribuição empírica e a teórica.

Aplicando o teste à variável resposta, considerando as distribuições Gama, Inversa Gaussiana e LogNormal, obtém-se os $p-values$ de 6.999e-07 e 2.2e-16 para as distribuições Gama e Inversa Gaussiana, respectivamente, o que indica a rejeição das hipóteses de que os dados provenham destas distribuições. Relativamente à distribuição LogNormal, obteve-se o $p-value$ de 0,3303, valor superior aos níveis de significância usuais, o que indica a não rejeição da hipótese de que a variável resposta se adequa à distribuição LogNormal. Esta tese é confirmada pelo gráfico da Figura 4.4, que avalia em que ponto a distribuição ajustada segue a distribuição cumulativa empírica. Logo, observa-se que a distribuição LogNormal se ajusta melhor aos dados em comparação com as distribuições Gama e Inversa Gaussiana.

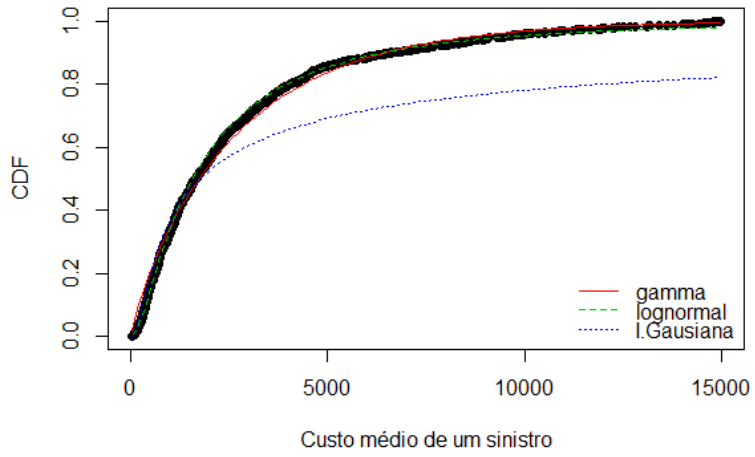


Figura 4.4: Gráfico Empírico CDF

4.3.2.2 Custo médio de um Grande Sinistro

Como referido na secção 4.3.2, procederemos à análise do custo médio de Grandes Sinistros, que surgem com frequência em carteiras de seguros automóvel, cujo custo de indemnização é bastante elevado.

A carteira apresenta apenas 73 apólices com custo de indemnização superior a 15 000 $, representando 5,28% do total de apólices indemnizados, tendo um custo médio por sinistro de 29 112,74 $ e o montante mínimo de indemnização é 15 015 $, como ilustra a Tabela 4.12. Nota-se ainda que, 99% dos grandes sinistros custaram no máximo 98 731,28 $, de acordo com a Tabela 4.13.

Tabela 4.12: Estatísticas de Custo Médio de Grandes Sinistros

Nº Apólices	Nº Apólice com custo	Nº Apólices com custo > 15.000	Custo médio	Desvio Padrão	Valor ($) Mínimo	Valor ($) Máximo
59 466	1 380	73	29 112,74 $	18 047,94 $	15 015	101 504

Tabela 4.13: Quantis dos Custos de Grande Sinistro ($)

50%	90%	95%	99%
22 249	41 731,60	63 423	98 731,28

Recorreu-se ao teste de ajustamento de Kolmogorov-Smirnov, com objetivo de determinar uma distribuição adequada para a variável aleatória. Considerando as distribuições Pareto, Weibull, Log-logistic e Pareto Generalizada, obtiveram-se p-values de 3.0811e-11, 0,001984, 0,2193 e 0,3743744, respectivamente. Assim, em função do p-value de cada teste, conclui-se que a variável aleatória pode ser modelada por uma distribuição Pareto Generalizada, por apresentar maior p-value e superior aos níveis de significância usuais.

4.3.3 Variáveis Explicativas

4.3.3.1 Exposição ao Risco

A variável tempo de exposição, é uma variável quantitativa, formada por valores fracionados calculados para cada uma das apólices. Com um tempo médio de exposição de 0.9783, que indica o tempo de exposição da maioria dos segurados é equivalente a um ano. Esta variável é considerada como um termo *offset* na modelação de frequência de sinistralidade.

4.3.3.2 Zona de Circulação (Províncias)

A variável Zona de Circulação é uma variável categórica codificada de 1 a 18 (que corresponde às 18 províncias de Angola) e segundo o gráfico da Figura 4.5, nota-se uma expressiva concentração dos segurados na Zona de Circulação 1 (que corresponde a província de Luanda), com a frequência de sinistralidade mais elevada. Comparando diferentes zonas em termos de similitude de frequência de sinistralidade e custo médio por sinistro, verifica-se existir alguma homogeneidade nalgumas zonas de circulação, tanto para frequência de sinistralidade, como para o custo médio por sinistro.

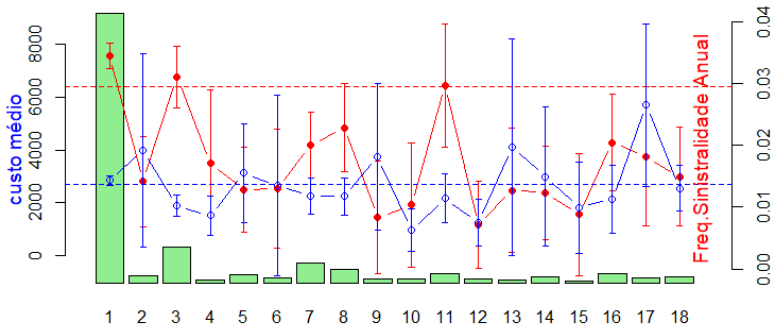


Figura 4.5: Frequência e Custo vs Zona de Circulação

4.3.3.3 Idade do Condutor

A variável Idade do Condutor é uma variável quantitativa discreta. A idade média na carteira é de 47 anos, em que o condutor mais novo possui 23 anos de idade, enquanto que o mais velho possui 99 anos de idade. Na Figura 4.6, observa-se que existe um elevado número de segurados entre os 46 e os 50 anos, e que os segurados com idade compreendidas entre os 81 e 99 anos, apresentam a frequência de sinistralidade mais alta e um dos custos mais elevados.

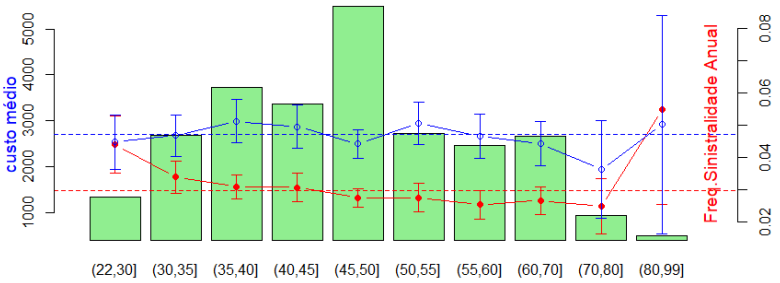


Figura 4.6: Frequência e Custo vs Idade do Condutor

Observando a Figura 4.6, nota-se um comportamento homogéneo relacionado com a frequência de sinistralidade em alguns intervalos etários, como é o caso do intervalo $(35, 40]$ e $(40, 45]$ e para custo médio por sinistro semelhantes encontramos os intervalos etários $(30, 35]$ e $(55, 60]$, bem como nos escalões $(22, 30]$ e $(60, 70]$.

4.3.3.4 Idade do Veículo

A Idade do Veículo é uma variável discreta quantitativa que corresponde ao número de anos do veículo. A média de idade dos veículos em carteira é de 9 anos e o veículo mais novo tem 3 anos de idade, enquanto que 58 anos é a idade do veículo mais antigo. Analisando a Figura 4.7, verifica-se que a maioria dos segurados possuem veículos com menos 9 anos de idade, com sinistralidade mais alta, enquanto que os veículos com mais de 14 anos apresentam a frequência de sinistralidade mais baixa e uma heterogeneidade entre os três intervalos etários, ou seja, apresentam frequência de sinistralidade e custo médio por sinistros bastantes dispares entre si.

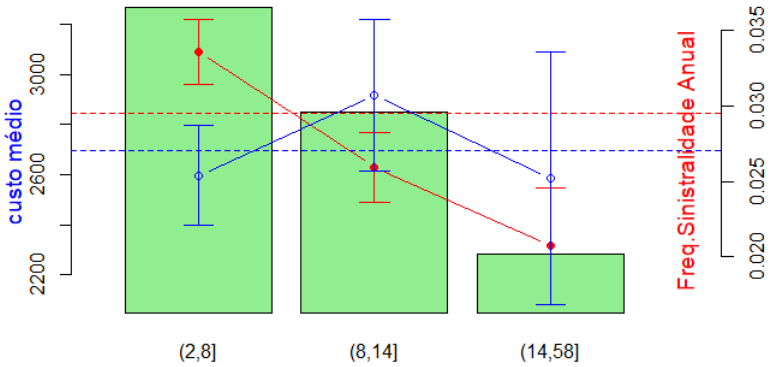


Figura 4.7: Frequência e Custo vs Idade do Veiculo

4.3.3.5 Anos de Carta

A variável Anos de Carta indica o número de anos que o tomador de seguro tem na condução de veículos a motor. A média do número de Anos de Carta nesta carteira é de 17,73 anos. Os anos de carta na carteira variam entre os 4 e os 69 anos. Segundo o gráfico da Figura 4.8, observa-se uma ligeira concentração dos segurados no intervalo entre 17 e 19 anos de carta, com a frequência de sinistralidade mais elevada e com custo médio por sinistro relativamente baixo. No que toca, à homogeneidade, nota-se a existência de intervalos homogéneos relacionados com a frequência de sinistralidade, o que não sucede no custo médio por sinistro.

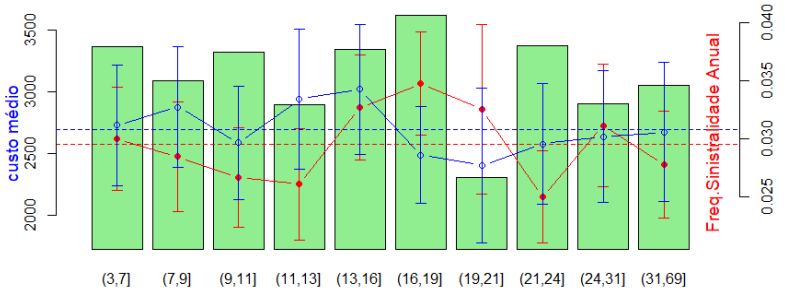


Figura 4.8: Frequência e Custo vs Anos de Carta

4.3.3.6 Número de Lugares do Veículo

Esta variável corresponde à capacidade do veículo em transportar pessoas. A carteira em análise comporta maioritariamente veículos de 5 lugares, conforme ilustra o gráfico da Figura 4.9, com a menor frequência de sinistralidade, enquanto que os veículos com mais de 9 lugares apresentam a maior frequência de sinistralidade e custo médio por sinistro.

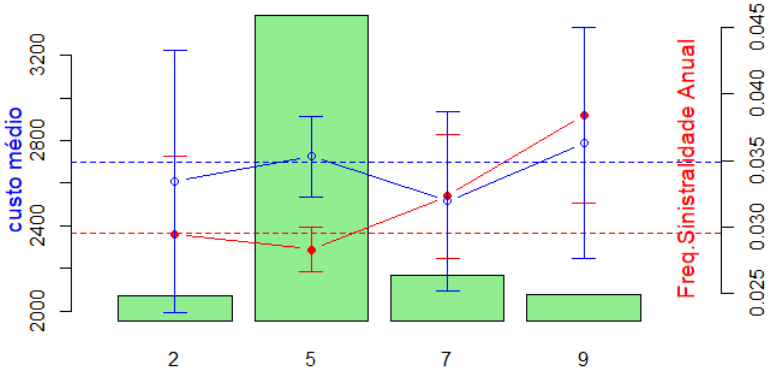


Figura 4.9: Frequência e Custo vs Número de Lugares do Veículo

4.3.3.7 Tipo de Uso

O Tipo de Uso é uma variável categórica codificado, que corresponde ao tipo de utilidade destinado ao veículo. Os veículos do tipo 111 (particular), representam 94,21% na carteira, com uma frequência de sinistralidade e custo médio por sinistro próximo da média global da carteira. Os veículos do tipo 112 (aluguer) apresentam

a frequência de sinistralidade mais elevada e um custo médio por sinistro a rondar o custo médio global da carteira. Segundo a Figura 4.10, nota-se homogeneidade entre a categoria 113 e 115 relativamente à frequência de sinistralidade, o que não sucede com custo médio por sinistro.

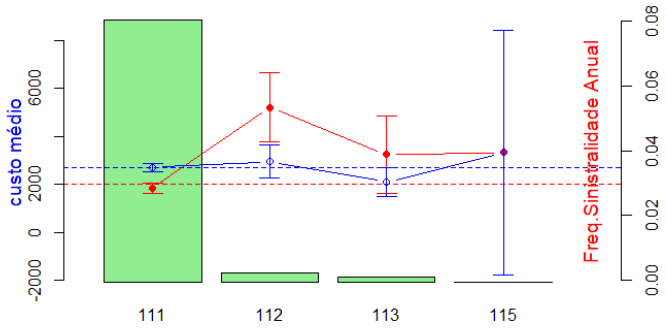


Figura 4.10: Frequência e Custo vs Tipo de Uso

4.3.3.8 Tipo de Veículo

O Tipo de Veículo é uma variável que corresponde à tipologia dos veículos que define a sua utilidade. Segundo a Figura 4.11, a carteira em análise possui mais veículos do tipo 1 (Ligeiro), com custo médio por sinistro em torno do custo médio global e os veículos do Tipo 2 (veículos pesados) apresentam maior frequência de sinistralidade, consequentemente maior custo médio por sinistros.

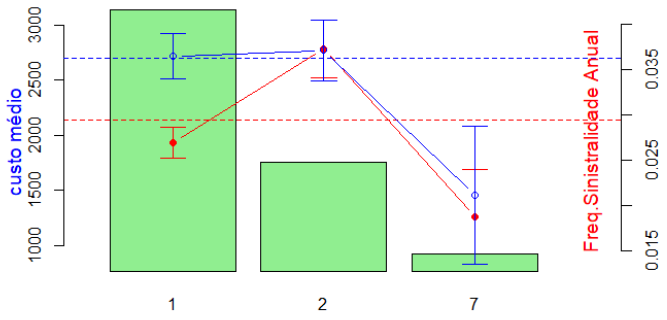


Figura 4.11: Frequência e Custo vs Tipo de Veículo

Apresenta-se na Tabela 4.14, todos os factores tarifários a utilizar na construção da estrutura tarifária, com os respectivos nível tarifários.

Tabela 4.14: Variáveis Tarifárias

Factor Tarifário	Nível Tarifário	Descrição	Factor Tarifário	Nível Tarifário	Descrição
Idade do Condutor	IC1	Dos 22 aos 30 anos	Zona de Circulação	ZC1	Luanda
	IC2	Dos 30 aos 35 anos		ZC2	Bié
	IC3	Dos 35 aos 40 anos		ZC3	Benguela
	IC4	Dos 40 aos 45 anos		ZC4	Bengo
	IC5	Dos 45 aos 50 anos		ZC5	Cabinda
	IC6	Dos 50 aos 55 anos		ZC6	Cunene
	IC7	Dos 55 aos 60 anos		ZC7	Huambo
	IC8	Dos 60 aos 70 anos		ZC8	Huíla
	IC9	Dos 70 aos 80 anos		ZC9	Cuando Cubango
	IC10	Dos 80 aos 99 anos		ZC10	Cuanza Norte
Idade do Veículo	IV1	Dos 2 aos 8 anos		ZC11	Cuanza Sul
	IV2	Dos 8 aos 14 anos		ZC12	Lunda Norte
	IV3	Dos 14 aos 58 anos		ZC13	Lunda Sul
Anos de Carta	AC1	Dos 3 aos 7 anos		ZC14	Malange
	AC2	Dos 7 aos 9 anos		ZC15	Móxico
	AC3	Dos 9 aos 11 anos		ZC16	Namibe
	AC4	Dos 11 aos 13 anos		ZC17	Uíge
	AC5	Dos 13 aos 16 anos		ZC18	Zaire
	AC6	Dos 16 aos 19 anos	Tipo de Veículo	TV1	Ligeiro
	AC7	Dos 19 aos 21 anos		TV2	Camioneta Camião, Pesado, Autocarro, Veículo Industrial e Misto
	AC8	Dos 21 aos 24 anos		TV3	Motociclos e Velocípedes
	AC9	Dos 24 aos 31 aos	Tipo de Uso	Uso1	Particular
	AC10	Dos 31 aos 69 anos		Uso2	Aluguer
Lugar do Veículo	LV1	2 Lugares		Uso3	Táxi
	LV2	5 Lugares		Uso4	Outros
	LV3	7 Lugares			
	LV4	Mais de 9 lugares			

4.3.4 Análise Multivariada

De modo a obtermos mais informações que nos facilitem a modelação, procedemos a uma análise multivariada que permite efectuar a correlação entre covariáveis ao mesmo tempo e extrair conclusões de relação entre elas. Considerando as variáveis

Número de Lugares do Veículo, Tipo de Veículo, Tipo de Uso, Zona de Circulação, Idade do Veículo, Anos de Carta, Idade do Condutor e observando a Figura 4.12, pode-se verificar a existência de variáveis correlacionados, sendo um factor a ter em conta aquando da modelação. Assim, verifica-se uma correlação moderada de 0.15, entre o Tipo de Veículo e Tipo de Uso. Como é de esperar observa-se ainda uma correlação razoável entre as variáveis Idade do Condutor e Anos de Carta e nota-se ainda uma correlação negativa entre Número de Lugares de Veículo e Tipo de Veículo. Portanto, não havendo variáveis fortemente correlacionados, proceder-se-á a modelação com todas as variáveis.

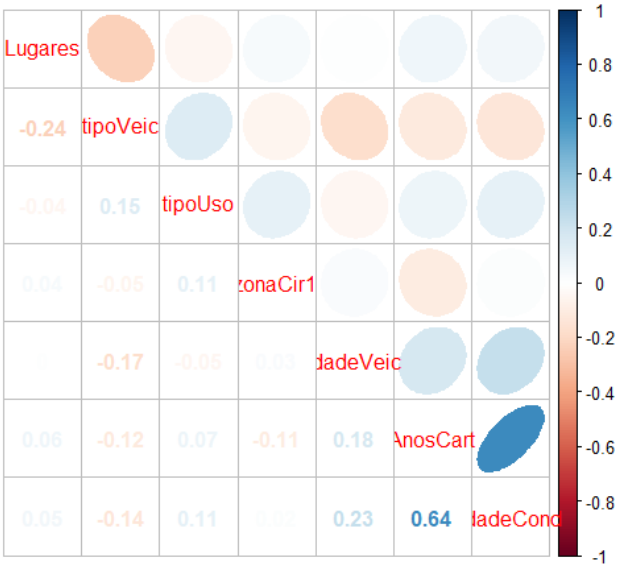


Figura 4.12: Correlação entre as Variáveis

A frequência média de sinistralidade para um segurado da carteira, tendo em conta a exposição de risco é de 0,0295, o que indica uma sinistralidade de aproximadamente 3% durante um ano. Cada um dos sinistros terá uma severidade, refletido no seu custo económico de 4 093,5 $, que corresponde ao custo médio de um sinistro.

O Prémio médio da carteira é de 120,8 $, que indica o montante económico que o segurado deve contribuir, em média, para manter o sistema equilibrado. No entanto, esta forma de calcular o prémio puro quebra um dos objectivos fundamentais dos métodos do sistema tarifação, que é ajustar o prémio ao risco que cada segurado comporta. Para este efeito, foram agrupados os factores tarifários em níveis tarifários, considerando a homogeneidade de risco existente, o que constituirá a construção do modelo tarifário.

4.4 Construção dos Modelos Tarifários

Nesta secção, efectuar-se-á a estimação dos modelos de frequência de sinistralidade e custo médio por sinistro, por meio dos Modelos Lineares Generalizados, que culminará na Construção da Estrutura Tarifária.

4.4.1 Tarifação *a priori*

A modelação das estruturas tarifárias pelos Modelos Lineares Generalizados, permite estimar o impacto dos factores tarifários sobre o valor esperado do Número e Custo de Sinistros permitindo, assim, a cobrança de um prémio adequado ao perfil de cada segurado. Para tanto, pretende-se realizar uma análise de consistência e significância estatística dos parâmetros estimados, além de uma avaliação crítica da qualidade do ajustamento e dos resíduos observados.

Os resultados descritos e discutidos daqui em diante, referentes à análise tarifária, obtidos por meio de uma sub-rotina de programação computacional desenvolvida no software estatístico *R Project*.

4.4.1.1 Características do Segurado Padrão

A Tabela 4.15, corresponde às características do Segurado Padrão de cada factor tarifário, constituído pelos níveis tarifários com maior número de segurados, de modo a garantir a robustez das estimativas.

Tabela 4.15: Características do Segurado Padrão

Factores Tarifários	Designação	Escalão
Idade do Condutor	$(45, 50]$	IC4
Idade do Veículo	$(2, 8]$	IV1
Anos de carta	$(16, 19]$	AC3
Zona de Circulação	Luanda	ZC1
Tipo de uso	Particular	Uso1
Tipo de Veículo	Ligeiro	TV1
Lugares de Veículo	5 lugares	LV1

4.4.1.2 Seleção do Modelo e Variáveis

O modelo de seleção de variáveis para Modelos Lineares Generalizados é efetuado pelo *Método Stepwise*, mediante o Critério de Informação de Akaike (AIC), mencionado

no capitulo 3 e também tendo em conta o valor de *p-value* do teste de nulidade de cada nível tarifário. Este procedimento proporcionará a seleção do modelo de melhor ajustamento.

4.4.1.3 Modelação de Frequência de Sinistralidade

Em consonância com a análise feita na secção anterior, proceder-se-á à modelação da frequência de sinistralidade, tendo em conta que a variável a modelar segue uma distribuição Binomial Negativa. Para tal, inicialmente procedemos à modelação individual das variáveis explicativas para apurar a significância dos níveis tarifários e, de seguida, a inclusão de todas as variáveis num único modelo, tendo-se obtido o resultado da Tabela 4.16, em que todos os factores são estatísticamente significativos.

Tabela 4.16: Modelação da Frequência de Sinistralidade em função das Covariáveis

	Estimate	Std. Error	z-value	Pr(>|z|)
Segurado Padrão	-3.32472	0.04422	-75.194	< 2e-16 ***
IC1	0.47510	0.10641	4.465	8.01e-06 ***
IC10	0.72740	0.27087	2.685	0.007245 **
IV2	-0.30228	0.05577	-5.420	5.97e-08 ***
IV3	-0.52211	0.09834	-5.309	1.10e-07 ***
AC4	-0.14372	0.06653	-2.160	0.030752 *
AC8	-0.21301	0.08961	-2.377	0.017453 *
ZC2	-0.97441	0.26451	-3.684	0.000230 ***
ZC5	-1.04812	0.26387	-3.972	7.12e-05 ***
ZC6	-1.14066	0.32275	-3.534	0.000409 ***
ZC7	-0.51009	0.14677	-3.475	0.000510 ***
ZC8	-0.44440	0.15573	-2.854	0.004321 **
ZC9	-1.56083	0.45161	-3.456	0.000548 ***
ZC10	-1.28948	0.41395	-3.115	0.001839 **
ZC12	-1.73602	0.50503	-3.437	0.000587 ***
ZC13	-1.13347	0.41533	-2.729	0.006351 **
ZC14	-1.13290	0.30726	-3.687	0.000227 ***
ZC15	-1.54980	0.58316	-2.658	0.007870 **
ZC16	-0.58654	0.19801	-2.962	0.003054 **
ZC17	-0.74330	0.28482	-2.610	0.009062 **
ZC18	-0.95163	0.27461	-3.465	0.000529 ***
Uso2	0.37556	0.11221	3.347	0.000817 ***
Uso3	0.49038	0.16335	3.002	0.002682 **
TV2	0.37351	0.05705	6.547	5.87e-11 ***
TV3	-0.70450	0.15138	-4.654	3.26e-06 ***

Analisando a Tabela 4.16, verificamos a existência de segurados com risco mais elevado que o segurado padrão, como é o caso dos segurados com mais de 80 anos de idade (IC10), que conduzem Veículo de Aluguer (Uso3) do tipo 2 (TV2). De modo contrário, encontramos os segurados que apresentam menos risco em comparação

com o segurado padrão, por exemplo, os segurados que conduzem veículos com mais de 14 anos de idade (IV3), possuidores da carta de condução com idade compreendida entre 21 e 24 anos (AC8), que conduzem motociclos e velocípedes (TV3) e que circulam frequentemente na Lunda Norte (ZC12).

4.4.1.4 Modelação da Severidade de Sinistros Regulares

Para a modelação de severidade de sinistro, considerando que variável resposta se ajusta melhor à distribuição LogNormal, como proposto na secção 4.3.2, procede-se de forma análogo a modelação de frequência de sinistralidade, modelando separadamente as variáveis tarifárias, com intuito de analisar a homogeneidade de custo médio por sinistro entre os níveis tarifários, de forma a agregá-las e/ou a incorporação de níveis não significativos ao segurado padrão.

Com a inclusão de todas variáveis tarifárias no modelo, utilizamos Teste de Razão de Verosimilhança para escolha do modelo, comparando modelo completo e restrito, de forma avaliar a influência das variáveis não significativas sobre o nível global do modelo, obtendo-se $p - value$ de 0,3766, que aponta uma evidência estatística de que não se deve rejeitar a hipótese de nulidade do modelo com quatro (4) factores tarifários ajustar melhor, em detrimento à hipótese do modelo com sete (7) variáveis (modelo completo). Assim, obtemos o resultado da Tabela 4.17, em que os factores tarifários são todas estatísticamente significativos.

Tabela 4.17: Modelação da Severidade de Sinistro em função das Covariáveis

	Estimate	Std. Error	t-value	Pr(>|t|)
Segurado padrão	7.44995	0.03285	226.784	< 2e-16 ***
IC9	-0.54517	0.21232	-2.568	0.01035 *
ZC3	-0.64553	0.09897	-6.523	9.87e-11 ***
ZC16	-0.58989	0.21249	-2.776	0.00558 **
ZC17	0.80390	0.34050	2.361	0.01838 *
TV3	-1.17802	0.23740	-4.962	7.89e-07 ***
LV1	0.41618	0.16836	2.472	0.01357 *

Com base nas estimativas dos coeficientes de regressão da Tabela 4.17, nota-se a existência dos segurados que proporcionam maior severidade de sinistros em relação ao segurado padrão, por exemplo, segurados que conduzem veículo com 2 lugares (LV1) e circulam frequentemente no Uíge, que implica um agravamento maior ao prémio base. Relativamente ao perfil de risco do segurado que origina menor severidade de sinistro em relação ao segurado padrão, encontramos segurados

com idade compreendida entre os 70 e 80 anos (IC9), que conduzem motociclos (TV3), cuja zona de circulação é Benguela (ZC3).

4.4.1.5 Modelação da Severidade de Grande Sinistro

Nesta secção vamos proceder a estimação do custo médio e a probabilidade de ocorrência de um "Grande" sinistro, considerando o exposto na secção 4.3.2. Para se obter a probabilidade de ocorrência de um Grande Sinistro, utilizamos a Regressão Logística, tendo-se obtido o resultado da Tabela 4.18.

Tabela 4.18: Modelo de Regressão Logística

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.000e+00	3.934e-16	2.542e+15	<2e-16 ***
IC2	-9.339e-18	2.830e-16	-3.300e-02	0.9739
IC3	5.127e-17	3.032e-16	1.690e-01	0.8666
IC4	-1.636e-16	3.096e-16	-5.280e-01	0.6004
IC5	-8.438e-17	3.404e-16	-2.480e-01	0.8056
IC6	2.915e-16	3.935e-16	7.410e-01	0.4634
IC7	1.968e-16	4.619e-16	4.260e-01	0.6725
IC8	2.372e-16	5.849e-16	4.060e-01	0.6874
IC9	2.171e-16	6.607e-16	3.290e-01	0.7443
IC10	1.650e-16	7.351e-16	2.250e-01	0.8236
IV2	5.731e-17	1.342e-16	4.270e-01	0.6718
IV3	1.678e-17	2.074e-16	8.100e-02	0.9360
AC2	4.865e-17	2.789e-16	1.740e-01	0.8625
AC3	-9.889e-18	2.692e-16	-3.700e-02	0.9709
AC4	-4.485e-19	2.732e-16	-2.000e-03	0.9987
AC5	-2.545e-17	2.438e-16	-1.040e-01	0.9174
AC6	-3.281e-17	3.641e-16	-9.000e-02	0.9287
AC7	-6.752e-16	3.417e-16	-1.976e+00	0.0556 .
AC8	-5.037e-17	3.759e-16	-1.340e-01	0.8941
AC9	-2.124e-16	3.830e-16	-5.550e-01	0.5825
AC10	-2.312e-16	5.255e-16	-4.400e-01	0.6625
ZC2	5.165e-17	2.660e-16	1.940e-01	0.8471
ZC3	3.027e-16	3.591e-16	8.430e-01	0.4046
ZC5	-5.153e-17	5.880e-16	-8.800e-02	0.9306
ZC6	-2.483e-16	5.152e-16	-4.820e-01	0.6327
ZC7	2.809e-17	5.138e-16	5.500e-02	0.9567
ZC8	1.604e-16	4.760e-16	3.370e-01	0.7380
ZC11	-3.202e-16	5.626e-16	-5.690e-01	0.5727
ZC16	2.349e-17	5.652e-16	4.200e-02	0.9671
ZC18	-2.918e-16	7.114e-16	-4.100e-01	0.6840
Uso2	1.905e-16	3.418e-16	5.570e-01	0.5807
Uso3	5.518e-17	3.740e-16	1.480e-01	0.8835
TV2	-4.051e-17	1.448e-16	-2.800e-01	0.7812
LV2	-4.679e-18	3.269e-16	-1.400e-02	0.9887
LV3	4.739e-17	3.772e-16	1.260e-01	0.9007
LV4	-1.406e-16	4.010e-16	-3.510e-01	0.7278

Analisando a Tabela 4.18, verifica-se que nenhum dos factores tarifários é estatisticamente significativo, para estimar a probabilidade de ocorrência de um grande sinistro, pelo que recorremos à proporção dos segurados que originam grande sinistros, obtendo uma estimativa de 0,05289855 de probabilidade de ocorrência de um grande sinistro, o que significa que se considerará que a probabilidade de ocorrência de uma grande sinistro é igual para todos segurados.

A estimativa do valor esperado do custo médio de um grande sinistro, através do Modelo Linear Generalizado, requer que a distribuição da variável aleatória pertença à Família Exponencial, o que não sucede, pois a variável resposta segue a distribuição Pareto Generalizada, como referido na secção 4.3.2.2. Assim, sendo G-Custo de um Grande Sinistro, e considerando que G $\sim$ DPG (7 279,4973; 0,3335), obteve-se $\mathbb{E}[G] = 25921,98$ \$.

4.4.2 Análise dos Resíduos

Com base no descrito na secção 4.4.1.5, que propõe analisar a qualidade do modelo Binomial Negativa (frequência de sinistralidade) e LogNormal (severidade de sinistro), através da análise de resíduos de Pearson e também recorremos ao gráfico QQ-plot, que verifica a normalidade dos resíduos, através da comparação dos quantis empíricos e os quantis da distribuição normal, na qual os pontos devem estar próximos da reta indicativo, ver [13] e [18].

Na análise de resíduos, para que o modelo seja bem ajustado é necessário que os resíduos estejam dispersos aleatoriamente em torno de 0, com variância constante, concentrados entre -2 e 2, segundo [6].

Portanto, analisando o gráfico da Figura 4.13, verifica-se que os resíduos estão dispersos em tornos de 0 entre -3 e 3 para o modelo Binomial Negativa, enquanto que para o modelo LogNormal, nota-se que os resíduos estão predominantemente entre -2 e 2 em torno de zero. Além disto, no gráfico da Figura 4.14, que corresponde ao gráfico QQ-Plot, percebe-se, que tanto para modelo Binomial Negativa, quanto para LogNormal, os resíduos se encontram alinhados de maneira satisfatória aos quantis teóricos com os quantis da distribuição normal, apresentando um nível aceitável de alinhamento ao longo da reta normalizada, apesar de um desnível no principio e no final; porém no geral, os modelo se adequam aos dados. Assim, com estas conclusões, estamos em condições para proceder o cálculo de prémio.

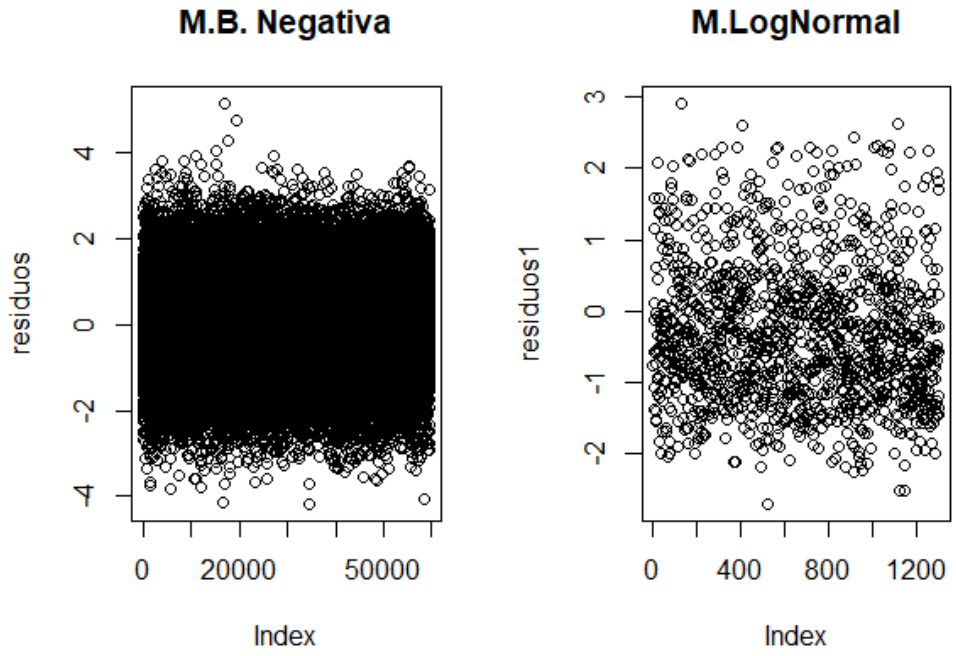


Figura 4.13: Resíduos

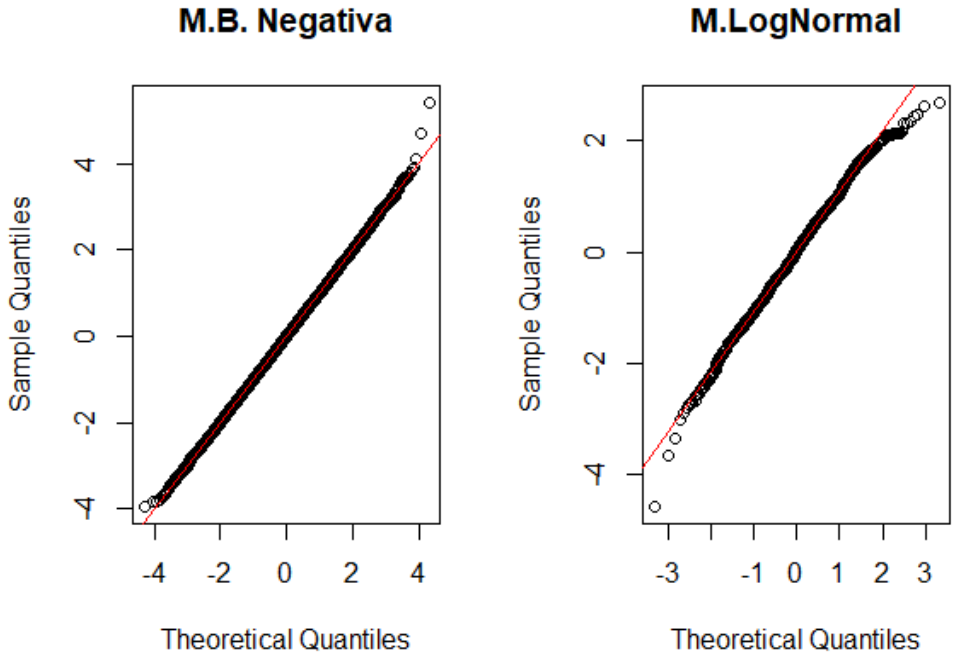


Figura 4.14: Gráfico QQ-plot

4.4.3 Cálculo do Prémio

Estimar os coeficientes de frequência e severidade de sinistralidade, e consequentemente o cálculo do prémio de seguro, constituem o principal objetivo deste trabalho. Portanto, após modelar os dados e estimar os elementos dos componentes dos modelos de frequência e severidade de sinistralidade, dado que o modelo considerado neste trabalho possui carácter multiplicativo e calcular as relatividades para o prémio consistirá em efectuar o produto entre as estimativas do modelo de frequência da sinistralidade, expressa pela

$$\mathbb{E}[N] = \mu_i = exp\left\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}\right\} \tag{4.2}$$

e as estimativas do modelo de severidade, dada pela

$$\mathbb{E}[Y] = \mu_i = exp\left\{\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + ... + \gamma_p X_{ip} + \frac{\sigma^2}{2}\right\} \tag{4.3}$$

Estas estimativas permitem mensurar o risco dos demais níveis de um determinado factor tarifário em relação ao Segurado Padrão (nível base de referência).

Na Tabela 4.19, apresentam-se os coeficientes estimados de frequência de sinistralidade e severidade de sinistro, incluindo os prémios associados a cada um dos escalões tarifários. Apresenta-se ainda a tarifa que indica a intensidade de agravamento ou desconto em relação ao Segurado Padrão, consoante o risco de ocorrência e gravidade dos sinistros. Salienta-se que na Tabela 4.19 estão incorporados os Grandes Sinistros, de acordo com a formula da equação (4.1) e referido na secção 4.3.2.

A partir da Tabela 4.19, podemos extrair características dos segurados que originam menor prémio, por exemplo, os segurados que conduzem motociclos (TV3) com mais de 14 anos de idade (IV3), detentores de carta de condução com idade compreendida entre 21 e 24 anos (AC8), que residem na zona da Lunda Norte (ZC12). Contrariamente, encontramos os segurados, que proporcionam maior prémio, como é o caso, dos indivíduos que conduzem veículo do tipo 2 (TV2), de 2 lugares (LV1), para efeito de táxi (Uso3), com idade compreendida entre 80 e 99 anos (IC10), que reside na zona do Moxico (ZC15).

A Tabela 4.20, apresenta o prémio do Segurado Padrão, bem como o prémio mínimo, que corresponde ao prémio que obteve mais desconto e prémio máximo, com mais agravamento do prémio base. A partir da Tabela 4.20, nota-se que a magnitude do prémio mínimo é demasiado baixo, tornando-se insustentável, de ponto de vista comercial. Portanto, deve-se adoptar um coeficiente de carga de segurança adequado,

Tabela 4.19: Estrutura Tarifária do Modelo Proposto

Factores Tarifários	Frequência de Sinistralidade	Severidade de sinistro	Prémio	Tarifa
Segurado Padrão	**0,0360**	**4258,121**	**153,207**	**153,207**
IC1	0,0579	4258,121	246,385	160,82%
IC9	0,0360	4337,618	156,068	101,87%
IC10	0,0745	4258,121	317,091	206,97%
IV2	0,0266	4258,121	113,240	73,91%
IV3	0,0213	4258,121	90,893	59,33%
AC4	0,0312	4258,121	132,697	86,61%
AC8	0,0291	4258,121	123,814	80,82%
ZC2	0,0136	4258,121	57,823	37,74%
ZC3	0,0360	4054,367	145,876	95,21%
ZC5	0,0126	4258,121	53,714	35,06%
ZC6	0,0115	4258,121	48,967	31,96%
ZC7	0,0216	4258,121	91,992	60,04%
ZC8	0,0231	4258,121	98,238	64,12%
ZC9	0,0076	4258,121	32,167	21,00%
ZC10	0,0360	4258,121	153,207	100,00%
ZC12	0,0063	4258,121	26,998	17,62%
ZC13	0,0116	4258,121	49,319	32,19%
ZC14	0,0116	4258,121	49,348	32,21%
ZC15	0,0614	4258,121	261,263	170,53%
ZC16	0,0200	4207,870	84,216	54,97%
ZC17	0,0171	12803,164	219,061	142,98%
ZC18	0,0139	4258,121	59,155	38,61%
Uso2	0,0524	4258,121	223,039	145,58%
Uso3	0,0588	4258,121	250,178	163,29%
TV2	0,0523	4258,121	222,584	145,28%
TV3	0,0178	2946,600	52,410	34,21%
LV1	0,0360	9129,007	328,462	214,39%

dependendo do principio de cálculo de prémio puro a ser adoptado, de modo que o produto seja rentável.

Tabela 4.20: Resumo dos Prémios ($)

Prémio mínimo	Segurado Padrão	Prémio máximo
4,428	153,207	2 306,011

4.4.4 Modelo Utilizado pela Seguradora

De acordo com informações da Seguradora, responsável pelos dados que suportam este trabalho, para o cálculo do prémio, considera-se que as variáveis aleatórias Número de Sinistros (frequência de sinistralidade) segue a distribuição Poisson e Custo por sinistro (severidade de sinistro) segue a distribuição Gama. Portanto, nesta secção, procederemos à modelação de frequência e severidade, segundo o modelo aplicado pela Seguradora, concomitantemente o cálculo de prémio e ainda faremos a comparação das Estruturas Tarifárias.

4.4.4.1 Modelação de Frequência de Sinistralidade

Para a seleção do modelo que apresenta melhor ajustamento, recorremos ao Critério de Informação de Akaike (AIC), tendo se obtido a Tabela 4.21, que corresponde ao valor de AIC para cada factores tarifários, cujos valores são superiores ao AIC do modelo, o que demonstra que todos factores são estatísticamente significativos para modelo. Assim, apresentamos na Tabela 4.22 o modelo final de frequência de sinistralidade e as estimativas dos parâmetros de regressão, cujos níveis tarifários são estatísticamente significativos.

Tabela 4.21: Seleção do Modelo de Frequência-AIC

Factores tarifários	Df	Deviance 12080	AIC 15447
Lugares de Veículo	2	12086	15449
Anos de Carta	2	12091	15454
Tipo de Uso	2	12092	15455
Idade do Condutor	3	12118	15479
Idade do Veículo	2	12128	15491
Tipo de Veículo	2	12173	15536
Zona de Circulação	15	12276	15614

Atendendo às estimativas dos coeficientes da regressão apresentados na Tabela 4.22, o perfil do segurado com maior risco é um indivíduo com mais de 80 anos de idade (IC10), que conduz um veículo com mais de 9 lugares (LV4), do tipo camioneta, camião, pesado, autocarro, veículo industrial e misto (TV2), para táxi (Uso3). Relativamente aos segurados com menor risco, encontramos indivíduos que conduzem motociclos e velocípedes (TV3) com idade compreendida entre 8 e 14 anos (IV2), possuidores da carta de condução entre 7 e 9 anos de idade (AC8) e residem na província da Lunda Norte (ZC12).

Tabela 4.22: Estrutura Tarifaria Final-Frequência de Sinistralidade

	Estimate	Std. Error	z-value	Pr(>|z|)
Segurado Padrão	-3.38326	0.04652	-72.734	< 2e-16 ***
IC1	0.53502	0.10182	5.255	1.48e-07 ***
IC2	0.26180	0.07855	3.333	0.000859 ***
IC10	0.74912	0.25292	2.962	0.003058 **
IV2	-0.29223	0.05422	-5.390	7.05e-08 ***
IV3	-0.50172	0.09600	-5.226	1.73e-07 ***
AC3	-0.17806	0.06575	-2.708	0.006765 **
AC8	-0.19562	0.08720	-2.243	0.024878 *
ZC2	-0.99623	0.26063	-3.822	0.000132 ***
ZC4	-0.86644	0.35491	-2.441	0.014634 *
ZC5	-1.05040	0.25986	-4.042	5.29e-05 ***
ZC6	-1.17021	0.31777	-3.683	0.000231 ***
ZC7	-0.51616	0.14314	-3.606	0.000311 ***
ZC8	-0.45136	0.15190	-2.971	0.002964 **
ZC9	-1.58449	0.44784	-3.538	0.000403 ***
ZC10	-1.31280	0.40929	-3.207	0.001339 **
ZC12	-1.74330	0.50025	-3.485	0.000492 ***
ZC13	-1.15302	0.40944	-2.816	0.004861 **
ZC14	-1.15237	0.30292	-3.804	0.000142 ***
ZC15	-1.59274	0.57799	-2.756	0.005857 **
ZC16	-0.57759	0.19259	-2.999	0.002708 **
ZC17	-0.75982	0.27890	-2.724	0.006443 **
ZC18	-0.95654	0.27025	-3.539	0.000401 ***
Uso2	0.29154	0.11158	2.613	0.008979 **
Uso3	0.40345	0.16202	2.490	0.012767 *
TV2	0.40816	0.05567	7.331	2.28e-13 ***
TV3	-0.72116	0.14869	-4.850	1.23e-06 ***
LV4	0.14338	0.06542	2.192	0.028407 *

4.4.4.2 Modelação de Severidade de Sinistro

Consideramos que a variável resposta do modelo segue uma distribuição Gama, sendo uma distribuição habitualmente utilizada para dados contínuos, com enviesamento à direita, admitindo valores bastantes elevados na cauda direita e frequentemente usada na modelação de dados de sinistralidade automóvel, mormente para custo por sinistros, tal como se observa em [14] e [22].

De forma análoga à secção 4.4.4.1, recorremos ao Critério de Informação de Akaike, para selecionar o melhor modelo para severidade de sinistro, cujos factores tarifários sejam estatísticamente significativos. Assim, a Tabela 4.23 apresenta valores de AIC para cada factor tarifário, em que se recomenda a exclusão dos factores com AIC menor ou igual 22 168.

Após a exclusão das variáveis tarifárias não significativas do modelo, obtemos na

Tabela 4.23: Seleção do Modelo de Severidade-AIC

Factores tarifários	Df	Deviance 1295.0	AIC 22168
Lugares de Veículo	3	1298.6	22165
Anos de Carta	1	1295.4	22166
Tipo de Uso	3	1296.8	22163
Idade do Condutor	2	1299.9	22168
Idade do Veículo	1	1297.3	22168
Tipo de Veículo	1	1310.0	22178
Zona de Circulação	3	1315.9	22179

Tabela 4.24, as estimativas dos parâmetros de regressão, para cada escalão tarifário, com respectivos $p-value$ de Teste de Wald, estatisticamente significativa, do melhor modelo para severidade de sinistro.

Tabela 4.24: Estrutura Tarifaria Final-Severidade de sinistro

	Estimate	Std. Error	t-value	Pr(>|t|)
Segurado padrão	7.94081	0.03252	244.162	< 2e-16 ***
ZC3	-0.38514	0.10059	-3.829	0.000135 ***
ZC10	-1.07361	0.54614	-1.966	0.049532 *
ZC17	0.70733	0.34633	2.042	0.041317 *
TV3	-0.64186	0.17531	-3.661	0.000261 ***

Tendo em conta os coeficientes da regressão indicados na Tabela 4.24, conclui-se que os segurados que residem na zona ZC17 (província do Uíge), apresentam maior severidade de sinistros em relação ao segurado padrão. Por outro lado, os segurados que residem na zona ZC10 (Cuanza Norte) e que conduzem motociclos e velocípedes (TV3), apresentam menor severidade de sinistro.

4.4.4.3 Análise dos Resíduos

Para o cálculo do prémio, é fundamental aferir a qualidade dos modelos encontrados. Para tal, recorremos à análise de resíduos, através da análise do comportamento dos gráficos Normais de probabilidades com envelopes simulados, que permitem verificar a adequação do modelo ajustado mesmo que os resíduos não tenham uma aproximação adequada com a distribuição normal, espera-se que para um modelo bem ajustado, que os pontos (resíduos) dispersos aleatoriamente estejam entre os limites do envelope, ver [23] e [29].

Observando o gráfico da Figura 4.15, verifica-se medidas de qualidade de ajuste e o comportamento dos resíduos no gráfico a total falta de aderência dos modelos. Esta tese confirma-se quando se analisa o gráfico QQ-Plot na Figura 4.16, onde se nota, tanto para o modelo de frequência de sinistralidade, como para a severidade dos sinistros, que os resíduos não estão alinhados satisfatoriamente com os quantis teóricos da distribuição Normal, apresentando um acentuado nível de desalinhamento com os pontos de dispersão ao longo da recta normalizada para a severidade, pouco menos para frequência.

Aconselha-se, desta forma, que a Seguradora reveja os pressupostos considerados no seu modelo de tarifação, de forma a melhorar adequação dos prémios ao risco da carteira.

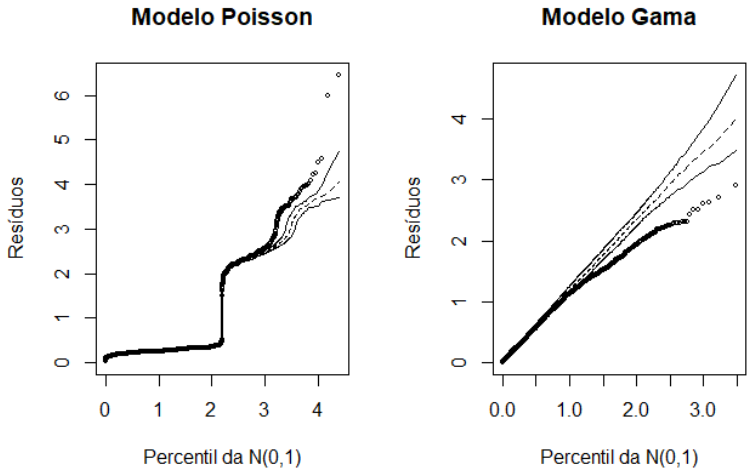


Figura 4.15: Gráficos Normal de Probabilidade com Envelope Simulado

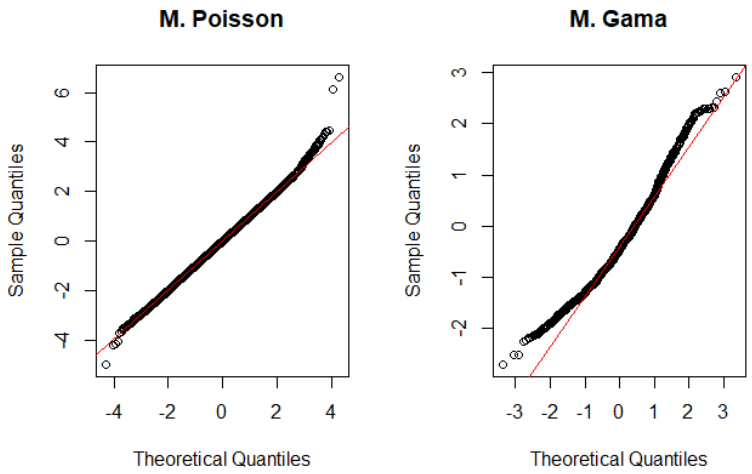


Figura 4.16: Gráfico QQ-Plot

4.4.4.4 Cálculo de Prémio

Com base nos modelos definidos de frequência de sinistralidade e severidade dos sinistro, procede-se à construção da estrutura tarifária. Na Tabela 4.25, encontramos a descrição das relatividades tarifárias associadas a cada um dos escalões tarifários, que exprimem em que intensidade o prémio será agravado ou descontado em relação ao segurado padrão. Referir que, a modelação da severidade de grande sinistro mereceu o mesmo procedimento efectuado no modelo proposta constante na secção 4.4.1.5.

A partir dos coeficientes estimados foi possível obter os prémios, efectuando o produto dos coeficientes de frequência e severidade tornando, assim, possível encontrar uma tarifa calculada para toda e qualquer combinação linear.

A partir da Tabela 4.25, encontramos as características dos segurados que apresentam agravamento e/ou desconto do prémio base. Para tanto, os segurados com idade superior a 80 anos (IC10), que conduzem camioneta, camião, autocarro, veículo pesado e misto (TV2), com mais 9 lugares (LV4), para táxi (Uso3) na zona do Moxico (ZC15), apresentam maior agravamento do prémio base. Já os segurados que conduzem motociclos e velocípedes (TV3), com mais de 14 anos de idade (IV3), portador de uma carta de condução com idade compreendida entre os 21 e 24 anos (AC8) e residem na zona do Lunda Norte (ZC12), possuem menor desconto do prémio base.

Na Tabela 4.26, apresentamos os segurados com mais e menos risco para a

Tabela 4.25: Estrutura Tarifária do Modelo da Seguradora

Factores Tarifários	Frequência de Sinistralidade	Severidade de sinistro	Prémio	Tarifa
Segurado Padrão	**0,0339**	**4032,240**	**136,841**	**136,841**
IC1	0,0579	4032,240	233,654	170,75%
IC2	0,0441	4032,240	177,793	129,93%
IC10	0,0718	4032,240	289,437	211,51%
IV2	0,0253	4032,240	102,165	74,66%
IV3	0,0205	4032,240	82,856	60,55%
AC3	0,0284	4032,240	114,521	83,69%
AC8	0,0279	4032,240	112,527	82,23%
ZC2	0,0125	4032,240	50,531	36,93%
ZC3	0,0339	3181,670	107,976	78,91%
ZC4	0,0143	4032,240	57,534	42,04%
ZC5	0,0119	4032,240	47,867	34,98%
ZC6	0,0105	4032,240	42,462	31,03%
ZC7	0,0203	4032,240	81,668	59,68%
ZC8	0,0216	4032,240	87,135	63,68%
ZC9	0,0070	4032,240	28,060	20,51%
ZC10	0,0339	2280,689	77,399	56,56%
ZC12	0,0059	4032,240	23,939	17,49%
ZC13	0,0107	4032,240	43,198	31,57%
ZC14	0,0107	4032,240	43,226	31,59%
ZC15	0,0614	4032,240	247,404	180,80%
ZC16	0,0190	4032,240	76,802	56,13%
ZC17	0,0159	6951,789	110,352	80,64%
ZC18	0,0130	4032,240	52,577	38,42%
Uso2	0,0454	4032,240	183,160	133,85%
Uso3	0,0508	4032,240	204,849	149,70%
TV2	0,0510	4032,240	205,816	150,41%
TV3	0,0165	2771,751	45,733	33,42%
LV4	0,0392	4032,240	157,938	115,42%

seguradora, tendo em conta o prémio do segurado padrão. Atendendo à magnitude bastante baixo do prémio mínimo, a Seguradora deverá aplicar um coeficiente de carga de segurança suficiente, de modo que o prémio seja equitativo e justa, para fazer face ao risco transferido.

Tabela 4.26: Resumo dos Prémios ($)

Prémio mínimo	Segurado Padrão	Prémio máximo
3,983	136,841	1 359,856

4.4.5 Comparação dos Prémios

Como referido no Capítulo 2, de que o prémio (puro) a pagar pelo segurado deve ser justo e equitativo, ou seja, deve ser proporcional ao risco transferido, caso contrário haverá prémio excessivo ou insuficiente, que poderá causar danos financeiro à Seguradora, bem como, originar eventuais perdas de carteira, se os prémios forem excessivos em comparação com o restante mercado Seguradora.

Analisando a Tabela 4.27, correspondente ao rácio dos prémios do modelo proposto e da seguradora, com base nas estruturas tarifárias das Tabelas 4.19 e 4.25, respectivamente, onde se verifica genericamente a sub-valorização do prémio cobrado pela Seguradora, ou seja, esta diferença significa que a Seguradora cobra (comercializa) o seguro a um valor inferior o que, no futuro, poderá tornar insolvente e/ou ineficiente a Seguradora perante o mercado. Por exemplo, um segurado com idade superior a 80 anos (IC10), que conduz camioneta, camião, pesado, autocarro, veículo industrial e misto (TV2), para Taxi (Uso3), reside no Cuanza Norte (ZC10), paga apenas 368,59 $, enquanto segundo o modelo aqui proposto, deveria pagar 752,26 $.

Tabela 4.27: Comparação dos Prémios

Factores Tarifários	Comparação dos Prémios
Segurado Padrão	1,1196
IC1	1,0545
IC2	0,8778
IC10	1,0955
IV2	1,1084
IV3	1,0970
AC3	1,1587
AC8	1,1003
ZC2	1,1443
ZC3	1,3510
ZC5	1,1222
ZC6	1,1532
ZC7	1,1264
ZC8	1,1274
ZC9	1,1464
ZC10	1,9794
ZC12	1,1278
ZC13	1,1417
ZC14	1,1416
ZC15	1,0560
ZC16	1,0965
ZC17	1,9851
ZC18	1,1251
Uso2	1,2177
Uso3	1,2213
TV2	1,0815
TV3	1,1460

Capítulo 5

Conclusões

O mercado de seguros em Angola encontra-se numa fase de crescimento, apesar da crise económica e financeira, o que pressupõe um grande desafio na comercialização dos produtos de seguros e um dos elementos fundamentais no negócio é a qualidade de serviço e o preço. Neste sentido, o cálculo do prémio (preço) merece total atenção a quando da sua concepção, pois à medida que o processo de tarifação incorpora um maior rigor técnico e de precisão ao cálculo do prémio, a companhia de seguros passa a ser cada vez mais capaz de cobrar um valor justo e que represente mais fidedignamente o risco associado ao perfil individual de cada segurado.

Das várias técnicas estatísticas empregadas para a realização de predição e a inferência sobre o comportamento de determinada variável aleatória de interesse, utilizamos os Modelos Lineares Generalizados para construir a Tarifa *a priori*, por ser um ferramenta predominante na análise tarifária, sustentado por uma base de dados de uma Seguradora Angolana.

Para tal, efectuou-se uma análise descritiva aos dados à cada variável tarifária, consequentemente a segmentação de grupos em riscos homogéneos. Adicionalmente, procedeu-se à modelação individual das variáveis tarifárias, posteriormente o encaixe num único modelo e consequente obtenção dos parâmetros estimados, o que permitiu analisar a influência, contribuições das variáveis tarifárias e o impacto dos parâmetros dos modelos sobre as variáveis aleatórias dependentes.

Além disso, observou-se que, de entre as distribuições de probabilidades adequadas para a modelação de frequência de sinistralidade e severidade de sinistros, consoante o p-value dos testes de ajustamentos de Qui-Quadrado e Kolmogorov-Smirnov, as que proporcionam melhor ajustamento aos dados foram distribuição Binomial Negativa para frequência, e distribuição LogNormal para severidade. Para efeitos comparativos considerou-se a distribuição Poisson para modelação da frequência e a distribuição Gama para modelação da severidade, num modelo utilizado pela Seguradora.

Através dos modelos estimados, procedeu-se à interpretação das estimativas de forma a compreender os efeitos destes sobre o comportamento esperado das variáveis aleatórias referentes à frequência de sinistralidade, severidade de sinistro e o prémio, analisando a intensidade dos coeficientes, indicando o agravamento e/ou desconto da tarifa e de que maneira as variáveis tarifárias e os respectivos escalões tarifários exercem influência na diferenciação do prémio a ser pago, tendo em conta o perfil de risco individual.

Procedeu-se ainda à comparação dos dois modelos (modelo proposto e o modelo em vigor na seguradora), tendo-se verificado discrepâncias significativas, o que justifica que a Seguradora deveria proceder a uma revisão tarifária.

Os resultados obtidos atendem ao objetivo deste trabalho, uma vez que foi apresentada uma metodologia completa para o cálculo do prémio de risco de um seguro de Responsabilidade Civil Automóvel, com base em dados reais.

Bibliografia

[1] J. Berkson. "Application of the logistic function to bioassay". Em: *American Statistical Association* (1944).

[2] M. Birch. "Maximum likelihood in three-way contingency tables". Em: *Royal Statistical Societyl* (1963).

[3] C. I. Bliss. *The calculation of the dosage-mortality curve*. Annals of Applied Biology, 1935.

[4] V. Brazauskas e A. Kleefeld. "Robust and Efficient Fitting of the Generalized Pareto Distribution with Actuarial Applications in View". Em: *University of Wisconsin-Milwaukee* (2009).

[5] A. Charpentier. *Computational Actuarial Science with R*. Montreal Canada: Chapman e Hall/CRC, 2014.

[6] G. Cordeiro e E. Neto. *Modelos Paramétricos*. Universidade Federal Rural de Pernambuco, 2006.

[7] J. V. Eeghen, E. Greup e J. Nijssen. *Rate making, Surveys of Actuarial Studies*. Vol. 2. National Nederlanden., 1983.

[8] "Exploring Heavy Tails Pareto and Generalized Pareto Distributions". Em: (2016). DOI: http://statmath.wu.ac.at/~hornik/QFS1/pareto-vignette.pdf.

[9] R. Fisher. *On the mathematical foundations of theoretical statistics*. Philosophical Transactions of the Royal Society, 1922.

[10] E. w. Frees. *Regression Modeling with Actuarial and Financial Aplications*. Cambridge University Press, 2010.

[11] E. w. Frees, R. A. Derring e G. Meyers. *Predictive Modeling Aplications in Actuarial Science*. Vol. II: case studies in insurance. Cambridge University Press, 2016.

[12] C. G. Giancaterino. "GLM, GNM and GAM approaches on TPML pricing". Em: *Mathematics and Statistical Science* (2014).

[13] *Gráfico Q-Q*. DOI: https://en.wikipedia.org/wiki/Q-Q_plot.

[14] G. R. Guerreiro. "Manual de Construção de Tarifa com R-O exemplo do seguro automóvel". Em: *FCT Nova* (2016).

[15] J. R. M. Hosking. "Parameter and Quantile Estimation for the Generalized Pareto Distribution". Em: *T. J. Watson Research Center* (1987).

[16] J. Kupper. "Some aspects of cumulative risk". Em: *International Actuarial Association* (1963).

[17] J. Lemaire. *Automobile insurance: actuarial models*. Springer Science e Business Media, 1995.

[18] J. K. Lindsey. *Applying Generalized Linear Models*. New York: Springer, 1997.

[19] M. de Lourdes Centeno. *Teoria do Risco na Actividade Seguradora*. Celta Editora, 2002.

[20] P. McCullagh e J. Nelder. *Generalized linear models (2nd edition)*. London: Chapman e Hall, 1989.

[21] Nelder e Wedderburn. *Generalized linear models*. Journal of the Royal Statistical Society, 1972.

[22] E. Ohlsson e B. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. 1 ed. Estocolmo: Springer, 2010.

[23] G. A. Paula. "Modelo de Regressão: com apoio computacional". Em: *IME-Universidade de São Paulo* (2010).

[24] S. T. dos Santos. "Construção de uma tarifação de Responsabilidade Civil". Universidade Nova de Lisboa, 2008.

[25] A. S. Sintra. "Métodos para Estimação dos Parâmetros da Distribuição de Pareto Generalizada: novas contribuições". Universidade de Lisboa, 2017.

[26] Turkman e Silva. *Modelos Lineares Generalizados*. Lisboa: Universidade de Lisboa, 2000.

[27] E. B. del Val. "Tarifation del Seguro del Automóvel: Métodos de Análisis Multivariante". Em: *Universidad de Barcelona* (2006).

[28] E. B. del Val, T. C. Cor e J. E. Fernandes. "Provisiones técnicas por año de calendario mediante modelo linear generalizado. Una aplicación con R Excel". Em: *Instituto de Actuarios Españoles* (2014).

[29] D. A. Williams. "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions". Em: *Royal Statistical Society* 36 (1987).

[30] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev e G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer Science, 2009.